**2024\_IEODataManagementTraining**

0:03
So. Welcome, everyone.

0:05
My name is Cat Pien.

0:06
I am a fish biologist at the Bureau of Reclamation.

0:10
And you'll be hearing my voice kind of walking through, presentations today.

0:16
I'm just going to kick us off, with a little bit of introduction of what

0:20
you're going to be hearing about today

0:21
and then framing our, our workshop.

0:24
So thanks to everyone for participating outside of the regular workshop slot.

0:29
We didn't really want to take away from, everyone's interesting science talks.

0:34
So, we decided to, to make this a little bit

0:39
this time and hope that it's valuable for everyone.

0:44
As Rosie mentioned, we will be recording the workshop,

0:47
so anyone who misses it can watch at a later date.

0:51
Or if you need to step out for for a period.

0:54
We'll send it out later.

0:58
But, we're

0:59
going to be framing our workshop around Open Science, which emphasizes

1:03
transparency along the entire scientific process

1:07
from data collection to analysis.

1:10
And the benefits of open science are to give you credit for the work you've done.

1:16
Get more citations.

1:18
Have people know what you're doing.

1:20
Has a bigger impact for your data.

1:26
Potentially saves you time from, providing data requests,

1:30
and so you can just provide a link and build trust with the community

1:34
by allowing them to see the data and the processes that you've been undergoing.

1:39
In recent years, the state

1:41
and federal government have been encouraging

1:44
or requiring an increase in making data more public and transparent,

1:49
which has led to there being more work on this front.

1:52
So I think part of part of the reason

1:54
we have these resources today are because those requirements

1:58
and the fair principles are often discussed in conjunction

2:03
with open science, as they describe some of the qualities desired for open science.

2:09
So, we're going

2:13
to have some polls today throughout our session.

2:16
And this is just kind of a kickoff poll.

2:18
So you get used to the process.

2:20
Some of you might have, used Menti before.

2:23
I think Rosie is going to put a link into the chat.

2:29
But the question is, do you know what FAIR

2:31
stands for?

2:49
So give people a minute.

2:53
Less than a minute.

2:57
And, the free version does only allow for 50 participants.

3:01
So first, 50 people response is recorded.

3:09
Okay.

3:09
So looks like.

3:14
About done.

3:16
Most of you got this right.

3:17
I think most of you have heard this acronym before.

3:23
Thanks, Dave.

3:26
Trying to get back.

3:29
okay. So,

3:32
FAIR stands for Findability

3:34
Accessibility, Interoperability and Reuse.

3:37
I'll quickly go through what each of these refers to.

3:41
So, findability

3:44
refers to data and metadata being findable or searchable.

3:48
often with an index.

3:50
So it's often associated with a globally unique and persistent,

3:55
so long lasting identifier.

3:58
So for example, on EDI where a lot of us post our data.

4:04
there's both an associated identifier and, digital

4:08
object identifier or DOI, which you often see with manuscripts.

4:12
that allows people to easily,

4:15
stick that into a web page and find your data,

4:19
even when it changes over time.

4:24
Accessibility

4:25
refers to being able to not only find, but also easily retrieve

4:30
and download the data by the identifier and also,

4:35
making your methods and protocols open, free and implementable.

4:39
So, methods that other people can access

4:43
and use.

4:48
Interoperability refers

4:51
to basically data being able to speak to each other.

4:55
So using shared and relevant vocabulary and terminology data standards.

5:01
so that data can be aggregated easily,

5:04
read easily combined with other sources.

5:07
So if you're using a certain kind of, for example, genetics data as

5:11
is cited, there's certain terminologies associated with that.

5:15
It also refers to data, metadata

5:18
being machine readable across different systems,

5:21
software and hardware.

5:26
And then Reuse.

5:28
refers to other people

5:30
being able to reuse your data and your methods.

5:33
So providing descriptive metadata, including

5:37
any complimentary files such as code and documentation

5:40
that are needed to understand the data and analysis.

5:44
also providing detailed information about any data sources.

5:49
You might not have collected your own data.

5:50
So, any data that you're using it,

5:55
information about the usage license.

6:01
And some of the challenges associated with open

6:04
data, open science are a lack of time and funding for people

6:09
to go through these processes, and a lack of knowledge for the proper,

6:13
proper or recommended procedures for sharing data.

6:17
and then, people

6:21
worrying about sharing data before it has been cleaned up.

6:25
Fear of misuse of data in other ways.

6:29
And then ownership of data and worry for, not having time to

6:34
to analyze your own data before putting it out things like that.

6:40
So today we're going to be mostly tackling number two.

6:43
So, we're hoping to provide knowledge for some of these open science processes,

6:49
but happy to talk about some of these other ones later on.

6:56
So, the IEP Data Utilization Workgroup is,

6:59
a group that has, it that's made up of a lot of the different,

7:04
member agencies of IEP, that come together and have been charged

7:09
with defining some of the shared data standards.

7:13
So, how to provide data documentation

7:16
descriptions across all member agencies.

7:20
And, the DUWG promotes efficient and open data management practices

7:24
by providing a lot of resources such as templates, guidance,

7:29
these kinds of trainings for the community.

7:33
The DUWG also provides support for member agencies in other ways.

7:38
Such as you can come to come to the group to get help publishing your data.

7:43
And then also coordinates across other groups to learn

7:46
some of the best, best practices and guidance that we learn from others.

7:51
Ssome

7:52
contact information below for reaching out to the,

7:56
organizers of the DUWG.

8:00
So for today, our objectives are to share the practices

8:04
and methods that support open science and fair principles

8:08
by providing you all with resources, templates,

8:11
and guidance for data management.

8:16
Sharing data sources, packages

8:18
and code for streamlined, transparent data analysis

8:21
and synthesis.

8:24
And connecting

8:25
the community with visualization tools, code, portals, and venues

8:29
for increasing and improving data communication.

8:37
For the agenda today in more detail,

8:40
we're going to start off the morning with data management and access topics.

8:46
Have a couple breaks in there

8:49
and the afternoon will be focused

8:51
on visualization related topics,

8:54
and finished off with a Q&A.

8:59
So, the first session will be a data management really quick overview

9:03
of, some of the topics in the data management lifecycle.

9:12
Coded access will be covering

9:14
how to use R to access online data

9:18
resources.

9:22
I have a longer presentation on metadata and data publication.

9:26
Both the importance of it and of how to, document and publish your data,

9:32
mostly focused around EDI, which is where most of us

9:35
are publishing our, cleaned up data.

9:42
And then the visualization session, we'll be covering

9:45
a couple fun tools for visualizing and sharing your data.

9:51
And some examples

9:52
of places that have pulled in data, and visualize

9:56
those provided queries, things like that,

10:00
for us all to use as SacPAS and Bay Delta live,

10:05
as well as some code for, a fun.

10:08
Here we go.

10:11
And then we're going to finish off with a panel Q&A.

10:13
So please save your, your, more detailed questions for the end.

10:17
Put any other questions in the team chat and we'll try to get to those

10:22
either, during talks or, during that Q&A session.

10:30
At the

10:31
end of today, we will be sharing out a resource list.

10:35
I think Rosie might put the link into the chat as well.

10:38
If you want to be following along, there'll be a resource list,

10:42
code and slides from this workshop

10:45
and a recording of the training.

10:51
And just a thanks to everyone who is participating.

10:55
everyone who has collected data,

10:58
everyone who, helped out with with the training.

11:02
And you can see the all the speaker bios, at this link

11:06
that will also get posted in the chat.

11:10
And that's all I've got for that intro.

11:13
I do have the next presentation also, so I am just going to hop into that.

11:21
Okay.

11:22
So really quick overview of all the steps of the data lifecycle.

11:27
This will be kind of, kicking off the start of the data management session.

11:35
So here's

11:36
a visualization of the data management lifecycle from U.S.

11:41
Fish and Wildlife Service.

11:42
And we'll briefly go through all of the steps included in this presentation.

11:48
Just wanted you to notice that the image indicates the process is iterative.

11:54
So you'll probably be going between these steps or all

11:58
around the cycle a couple times, during your project.

12:03
And the QA/QC is in the middle part of every step of the process.

12:11
We're going to start with the plan stage.

12:13
This refers to data management planning, or creating a data management plan,

12:19
which describes how your data are collected

12:22
and administered over its entire lifecycle.

12:25
There are a lot of documents you'll hear about today

12:28
that include, topics similar to what is in the data management plan.

12:34
I'll show you, some of the, the fields that can be in there

12:40
on the next slide.

12:41
The DMP is meant to be a short communication piece focused on,

12:47
all of your data processes, and it'll accompany

12:51
other types of documentation, such as SOPs, metadata, QAPPs

12:56
that provide a lot more detail about things like your collection

12:59
methods, your quick procedures, historical changes and things like that.

13:05
It is a living

13:06
document to be continuously updated as your processes change,

13:10
and can be applicable to anything from a long term monitoring study

13:14
to short field study or analysis, or

13:17
any other kind of study.

13:21
So just a quick,

13:23
view of some of the, topics you'll want to include in your data

13:28
management plan. We'll provide resources for example ones and templates.

13:34
So, just wanted to show you a quick preview.

13:40
So the next stage we're going to is the acquire stage.

13:44
This is, referring to the collection of new data.

13:48
So, you know, going out in the field collecting data or,

13:52
using and transforming existing data for,

13:56
you know, synthesis, analysis projects

13:58
or to help supplement your, your data collection.

14:04
The maintain

14:05
stage refers to basically the maintenance of your data.

14:09
So, figuring out what database you're using

14:12
and managing that database, checking your data.

14:15
So QAQC

14:18
recording and loading your data after you've collected it.

14:22
and then all the processes after, in your database.

14:26
So things like renaming, combining data, converting it into a usable,

14:30
shareable format.

14:35
Maintain also

14:35
refers to documenting your data.

14:39
So, describing your equipment, your field and lab procedures

14:44
parameters and, documenting any changes

14:49
that you make to those procedures, as well as your version changes.

14:54
You'll hear more about metadata later.

15:00
Access refers to providing access for your data,

15:04
whether it is, providing

15:07
a web version of your database,

15:10
publishing your data on EDI or another data portal.

15:15
putting your data on a visualization portal like SacPAS or Bay Delta Live,

15:19
or publishing your data in a data paper, report or manuscript.

15:30
And then evaluate.

15:31
This is where you get to start looking at the trends,

15:36
evaluating, analyzing, plotting your data.

15:40
In this process as well, developing reproducible workflows

15:44
so that others can replicate these analyses

15:48
and processing that you're doing.

15:53
Then,

15:54
archive last step before going back to plan.

15:58
so this is where you are archiving and securing your data.

16:03
For example, making multiple copies

16:06
of your data database in case something accidentally gets deleted.

16:11
Scanning your data sheets, if you have physical data sheets. Storing everything

16:16
in a shared and backed up location, and securing

16:19
any confidential data.

16:24
And then for quality assurance and control, as I mentioned,

16:28
these processes are relevant along the entire process.

16:33
you're going to hear more about this in the next presentation.

16:36
So just, a brief, definition for quality assurance is

16:41
this is where you're ensuring data quality on the front end of data collection,

16:45
such as training and writing SOPs

16:50
and quality

16:51
control refers to the data quality post collection.

16:55
So like outlier detection, removing,

17:00
questionable data.

17:02
And that's where we're going to go next.

17:05
So, I'm going to pass this over to Rachel

17:10
to talk more about quality assurance and quality control.

17:17
All right.

17:18
let me go ahead and share my screen here.

17:25
Okay, can everybody see my screen now?

17:31
Yes. Okay, great.

17:32
Thank you.

17:33
All right. So. Yes. Hi.

17:34
My name is Rachel Pisor, and I'm a senior Environmental scientist

17:37
supervisor with the DWR quality assurance program.

17:41
Today I'll be talking to you about some things to consider with quality

17:44
assurance and quality control of your projects.

17:49
All right.

17:50
So first let's get on the same page with what these terms mean.

17:53
So QA and QC our terms that are often used interchangeably

17:57
because they're tightly linked concepts.

17:59
But they are two different things.

18:00
A quality assurance is an integrated system involving planning,

18:04
implementation, assessment, reporting and continuous improvement.

18:08
It involves planning to ensure that the quality data will be achieved

18:12
and uses processes, standards, procedures and or policies to accomplish it.

18:17
The goal is to prevent or eliminate errors rather than to identify

18:21
and then fix them.

18:22
Typical components include quality assurance project plans, standard

18:26
operating procedures, and training.

18:28
QA starts at the concept of a project and follows through

18:30
to the archiving of data and publishing of the final report.

18:34
Quality control is a component of quality assurance,

18:37
and it is about determining if the quality of the data meets the requirements,

18:41
is compliant, and conforms to the standards that you determined.

18:44
This happens you know when you're in your planning phase.

18:48
It involves various kinds of checks of the data

18:53
before the data is deemed final and ready for analysis or publishing.

18:59
So QA starts with the

19:00
planning stages of the project, such as with the project plan or QAPP.

19:05
QAPPs include or reference many of the supporting planning documents listed here.

19:11
Data management plans are also a common tool.

19:13
They typically have a lot of information regarding the data management

19:15
and can be referenced in the QAPP, but the QAPP will be going

19:20
into a lot more detail.

19:21
QA for data management answers questions like how the data

19:26
is entered into a database, what kind of database it is stored in.

19:30
How are the project documents being organized, backed up and archived,

19:34
and how field sheets and records are going to be managed and archived?

19:37
Also, how would the laboratory data, if any, be stored?

19:41
Where will the project records like field sheets be maintained,

19:44
preserved and accessed?

19:48
So here are some examples of the types of documents and checks

19:51
you may be using to ensure quality control of your data.

19:55
There will be certain activities that are consistent

19:57
no matter what type of data you're collecting.

19:59
The components listed here, would likely be applicable

20:03
as long as you're collecting some kind of field data, be it

20:06
water quality, environmental or biological.

20:11
Each data

20:12
type may have its own QC checks to consider.

20:15
For discreet water quality data,

20:18
This could involve

20:19
checking your blanks to ensure there was no contamination in your sample.

20:23
Checking laboratory QC to make sure the laboratory processes

20:27
were in control, and checking constituent pairs

20:30
such as a total constituent having a higher concentration

20:33
than its dissolved analog. For real time water quality,

20:37
this could involve checking your pre and post calibration data

20:40
and evaluating the calibration and fouling error. For biological data,

20:45
QC could include having a secondary person confirm the ID

20:50
of the of the species or measurements like fork length.

20:54
For samples such as for fin clips or tissue,

20:57
it's verifying that the samples were collected

20:59
following standard procedure and are properly labeled and preserved.

21:03
For environmental data,

21:05
it could be having consensus with a second person on cloud cover,

21:09
for example.

21:13
There are various QC checks that are done

21:15
throughout the data management lifecycle, from initial collection to archiving.

21:20
Listed here are some examples, such as a completeness check

21:23
that the field sheets are filled completely, correctly, and legibly

21:26
if you're handwriting your data, and that if you have no data

21:30
for certain fields that is handled properly consistently.

21:34
The next step is hand entering

21:36
data or digitally uploading the data into a database.

21:39
Then that data is verified either through a second person, double checking

21:43
that the data in the database matches the field sheets exactly,

21:47
or another option is through double data entry.

21:50
This process involves having two staff independently enter

21:55
the data into the database, and then the duplicate entries are compared.

22:00
Any entries that are not identical are identified for further verification.

22:05
Electronic data checks are preliminary checks

22:07
that could be done with software such as Excel, Access, or R.

22:11
For example, making sure that the data values are

22:15
within the known possible range.

22:19
QC of the data may also involve data visualization

22:22
to identify potential outliers or unusual trends.

22:26
Visualization may also pull in different constituents of the data to compare,

22:30
or by looking at different time scales, such as this example

22:34
of a heat map showing the calibration error across a program's

22:37
stations, over the course of a year.

22:41
There are also various statistical tools

22:43
that can be used for identifying outliers for further examination,

22:47
based on what kind and frequency of data you're looking at.

22:51
There are a couple of resources I'm going to mention in the next couple of slides.

22:54
I can give you some ideas.

22:57
When doing all this QC,

22:59
it is important to qualify the data appropriately when it does not pass,

23:04
these specific checks, like through flagging.

23:08
This is critically important for transparency and context of the data

23:11
for those downstream users or any users really.

23:15
So further steps and the QC of the data management side is ensuring project

23:20
records and files are being appropriately named, stored, backed up and archived.

23:27
And so now I just wanted to share some of the resources available.

23:31
So, on the IEP DUWG website, there are some best practices documents

23:36
focused on fish QC and use of digital data sheets.

23:40
And there are also templates on the standard operating

23:42
procedures and metadata.

23:44
And of course the data management plans.

23:47
And then on the DWR outlier detection working group

23:51
GitHub, there is an R package with tasks associated with outlier detection.

23:55
So that that's worth checking out.

23:58
And then we also have some

23:59
DWP resources that are available on this box account.

24:03
This box link rather.

24:05
So, there is a QAPP template.

24:07
It is, it is focused on discreet water quality, but it,

24:11
it is easy to translate to

24:15
or apply to other data types as well.

24:19
There's best practices documents as well on outlier detection,

24:23
discreet water quality data review

24:26
and lab QC data review.

24:30
And with that, yeah, just reach out

24:33
if you have any questions regarding QA or anything on this presentation.

24:37
And I'll now stop sharing and pass it back to Cat.

24:42
Thanks, Rachel.

24:45
Thank you.

24:45
So next, I think I was actually supposed

24:50
to go to Rosie before Rachel, but next we'll have Rosie Hartman,

24:53
from DWR, talk a little bit about data structure and format.

24:58
Thanks, Cat.

24:59
I was wondering if you'd forgotten about me.

25:01
All right, so I'm going to talk a little bit about data structures and formats,

25:06
hopefully not too quickly because,

25:09
last minute I realized I have less time than I thought I did.

25:13
But first off, I want to say that everything I'm going to talk about today

25:17
in terms of good ways to structure your data, are some best practices

25:21
the DUWG and other data scientists, adhere to.

25:24
But they're not firm, fast rules.

25:27
There's definitely situations

25:28
where you might want to structure your database differently, but,

25:32
one of the first steps you do when you collect data,

25:36
is you need to figure out what to do with it because,

25:40
having a bunch of data sheets in a shoe box isn't really a data management plan.

25:44
and you probably have a really nice data sheet

25:47
that's laid out to be easy to fill out in the field.

25:50
This is a beautiful example from the Yolo Bypass Fish Monitoring Program.

25:54
It's easy to fill out, even in the rain.

25:58
I know from personal experience,

26:00
but we don't necessarily want to store our data in this format.

26:03
For one thing, we have three different types of data on the sheet.

26:07
We've got information about the surrounding environment.

26:10
We've got information about sampling effort, which we might have

26:13
multiple samples at one site with similar environmental information.

26:19
And then we have, fish.

26:20
And we'll probably have a lot of fish for each sample.

26:25
So how do we put this in a database.

26:27
We want a database that is relatively consistent.

26:31
We don't want to be making changes every six months.

26:33
So we want to, make it flexible enough to take in any kind of,

26:39
data that we're likely to collect in the future.

26:42
We want it to be efficient, without a lot of repeated rows,

26:46
and able to quickly enter data, as well as quickly pull data out of it.

26:51
We want it to be as simple as possible.

26:53
because if we have a ton of interrelated tables

26:58
and complicated queries and complicated forms, more things are likely to break.

27:03
It'll be difficult for people in the future who didn't

27:06
create it to maintain it.

27:08
And that gets into sustainability, especially if you have a third party

27:13
build a database for you.

27:14
You want to make sure that their training

27:16
your staff to use it and maintain it over time.

27:21
A few basic rules.

27:24
These are as close to rules as we get.

27:29
but as you're setting up

27:30
your database, you want to have each variable in its own

27:34
column, each observation, its own row, and each value in its own cell.

27:40
This is sometimes known as tidy data formats or, database normalization.

27:45
And and sometimes this will require your database to have multiple tables

27:50
instead of just one big table.

27:53
What does this really mean?

27:56
Well, I think the easiest way to look at this is to find out what not to do.

28:00
So I'd like you guys to humor me.

28:03
And just in the chat, write down anything you see with this table

28:07
that looks not that great.

28:11
Anything you see wrong with this table?

28:16
Give you 30s or so.

28:17
Here.

28:28
Come on.

28:29
There's 79 people.

28:31
There we go.

28:31
There we go.

28:37
Yeah.

28:38
Okay, so a lot of people picked up,

28:41
most of the things that I was looking for.

28:44
We have repeated information.

28:50
So the site information is repeated.

28:53
We have multiple columns with the same type of info.

28:55
Species

28:56
one, species two. Count and length.

29:00
We have comments in numeric columns.

29:05
That's going to give all sorts of problems.

29:07
Multiple values in one cell.

29:09
Very bad idea.

29:10
Multiple columns with the same name.

29:12
Even worse idea a lot of programs won't even let you do that.

29:16
Multiple indicators are missing values.

29:19
That's problematic.

29:20
And then you've got valuable info in the comments column.

29:25
So what should you do?

29:26
Well, the sort of best practices

29:29
is to take those, three types of information for your data sheet,

29:33
break them into separate tables so you stop the repeated rows.

29:37
We'll have one for that surrounding and for environment information,

29:41
one for the sample info.

29:43
And we'll have several samples per site visit

29:47
and then one for, the information about each fish.

29:52
But, how are these things working together?

29:56
Right now, we have

29:59
each column has a single type of information.

30:02
Each row is a single observation.

30:04
Each cell has one value.

30:06
But now we have like date and tide being separate from fish.

30:10
What do we need to do to connect these?

30:12
Well, the key is keys.

30:16
Each table has a unique feel a field indicating unique values.

30:20
This is called the primary key.

30:22
You can just have an arbitrary auto number key

30:27
or an informative key like date,

30:30
station, time, something like that.

30:34
But the important thing

30:36
is it cannot repeat or crazy stuff happens.

30:39
Those keys are what's used to link tables.

30:42
For example, these three tables,

30:44
the first column in each table is your primary key.

30:49
Visit id sample d fish ID.

30:53
You'll notice that those columns are repeated in the table below it.

30:57
Visit ID here.

30:59
Get my laser pointer visit already here.

31:01
Goes to visit ID here.

31:02
And it is repeated in this table.

31:04
That is what we can use to link the tables together. Same with sample ID

31:09
here goes to sample ID here.

31:12
This is a little easier to see on what's called a linkage map.

31:17
Access does this for you automatically.

31:20
And you see that visit ID from the site visit

31:23
table is linked to the sample info table and it's a one to many relationship.

31:27
So it'll be one value a visit ID and that value will be repeated multiple times.

31:32
And sample info.

31:33
Same with these two tables.

31:36
You might also end up with many to one relationships.

31:39
We have a lookup table, and one value of species code

31:43
and many values of species code in fish info.

31:47
If you end up with a many to many relationship,

31:50
there's probably something wrong because your data will balloon in size and

31:55
it's bad.

31:56
So if you end up with a manage many relationship, double check what you did.

32:01
So these are the linkages.

32:03
And you use those linkages to pull data out in the format you want.

32:08
You join together tables

32:10
you want. Then select columns you want, and filter rows you want.

32:15
And those are the terms used in SQL,

32:19
which is the language commonly used in database management,

32:22
but very similar syntax is available in R as well.

32:28
When you join tables on

32:29
that primary key, you end up with, you know, two tables.

32:33
These each have two columns.

32:35
The final table has three columns.

32:38
The key is what they joined on.

32:40
So that just gets merged.

32:43
And then you have all of the remaining columns

32:47
As an example, to join our site

32:49
if we wanted to get, say, the number of Chinook salmon we caught by day.

32:55
Date is in our site visit table,

32:59
and salmon is in our fish table.

33:01
So we'll start by joining the site visit to the sample info table

33:05
by visit ID. We end up with repeated site information, which is bad when storing it.

33:11
But in this case, it's okay because we need to use it

33:13
this way.

33:15
And we have all of the columns

33:16
from site visit and all the columns from sample info.

33:20
We can then join

33:22
the fish info table to our new table

33:25
based on sample ID. We have sample ID

33:29
in here, sample ID in here.

33:32
We join and now we have all of the columns from all of the tables.

33:37
Some of those columns are probably not that useful.

33:40
Our primary key is for example now that things are joined together

33:43
we might not need them. We could drop those.

33:46
We also might not be interested

33:49
in the fork length if we just want the total number of salmon.

33:54
So we can select the columns we want,

33:57
and then we can filter just the rows

33:59
that have species code of Chinook salmon.

34:03
And we get the number of Chinook salmon by day.

34:07
Sso that's kind of a little bit about

34:12
filters, queries.

34:14
There's lots, lots more that you can do with queries.

34:16
And it's pretty agnostic as to what platform you're using.

34:22
All databases have the same basic idea behind,

34:25
querying, joining, filtering and selecting.

34:30
But back to setting up your database.

34:33
What do you call things?

34:34
Well, it's a good idea to keep your column names short and informative.

34:38
Avoid any spaces or special characters that'll make it difficult

34:41
to read into another program later.

34:44
You also don't want to have the same column names and multiple tables

34:47
if they actually mean different things like sample date

34:51
and lab analysis date. Call them sample date,

34:54
lab date, not just date

34:56
in both. Definitely don't have the same,

34:59
don't have multiple columns of the same table, the same name.

35:02
Most, database programs won't even let you.

35:05
So, Yeah.

35:08
For example, sample date is good.

35:10
Date might not be so good.

35:12
Distance underscore meters versus distance based parentheses.

35:15
We go this is going to give you problems.

35:18
Bio volume better than the longer version etc..

35:21
And IEP put together a document of naming conventions.

35:24
So if you're picking names for columns we have some suggestions for you.

35:29
That's available online.

35:32
Missing values.

35:34
You should always use the comment field to say why a value is missing.

35:38
Don't put it in the value field.

35:40
And you should use a single consistent indicator for missing values.

35:44
You can just leave a blank space, which is often the easiest.

35:49
There is potential for confusion if you don't have a good comment as to

35:52
whether just forgot to enter the data, or the data is actually missing.

35:56
NA is kind of the gold standard

35:59
that everyone uses in data science for missing value.

36:02
But if you have a numeric column and you try and put it

36:05
in a character in a, sometimes it doesn't work.

36:08
Some people also use, a negative

36:10
999 or some other ridiculous looking value.

36:15
I don't

36:15
actually recommend that because if your data user isn't paying

36:19
attention, they'll just take all your data, calculate the mean,

36:23
and they're like, wait, why did we get -7000 fish?

36:26
That doesn't make any sense.

36:28
So not my favorite,

36:31
but it's possible.

36:34
The comment field should be for why, not

36:37
what. If you make the same comment again and again

36:40
That should be its own column.

36:42
So, something like the secchi was broken.

36:44
That explains why we have a missing value.

36:46
Or fish missing its tail.

36:48
Cow poop in the water could have been causing that really low DO value.

36:51
So, we know that the DO value isn't a mistake.

36:55
Now, if you see cow poop a lot, maybe you need a cow poop column.

36:58
But, in general,

37:01
if you say dead fish,

37:02
you're probably gonna have a lot of dead fish that might need to be its own column.

37:05
Something like a genetic sample. ID.

37:08
If you're doing this on a regular basis, that should also be its own column.

37:12
But if it's a special study that you're just doing once in a while,

37:14
then, yeah, put it in the columns, comment.

37:19
I'm not going to tell you

37:21
how to build this database, but, a few options.

37:26
Microsoft Access, if you just have a small amount of data

37:30
and are not a super computer nerd, that's the best way to go.

37:35
It's easy to use, easy to set up, user friendly,

37:41
and has a lot of advanced functionality.

37:43
There are rumors that Microsoft is going to start

37:45
stop supporting Access, but that hasn't happened yet,

37:47
and there's no equivalent easy database program that I've seen.

37:52
ESRI products for geo databases

37:54
if your data has a strong spatial component, that's a good option.

37:58
More advanced options are SQL server and custom built databases.

38:02
You can just have a lot of flat files that are linked based on primary keys,

38:08
but there is the potential for things to get messed up there.

38:10
So not my favorite.

38:13
Some further

38:15
reading in the, resources and I'm over time,

38:18
so I will pass it on to, the next person.

38:23
Thank you very much.

38:26
Thanks, Rosie.

38:27
Yeah, we can always shift a little.

38:29
Maybe shave a couple minutes here and there off

38:31
lunch or break.

38:33
But let's pass it on next to Denise from U.S.

38:38
Fish and Wildlife Service.

38:43
Hey, hello.

38:44
I'm Denise Goodman with the U.S.

38:47
Fish and Wildlife Service in Lodi.

38:48
And I'm going to be talking about data sharing by the Enhanced

38:51
Delta Smelt Monitoring Program, or EDSM.

38:54
EDSM, is a delta smelt sampling program.

38:58
We sample Delta smelt through the majority of their lifecycle.

39:02
But one of our major objectives is to support

39:04
real time water operations and management decisions.

39:09
This was a huge point of emphasis when EDSM was first established in 2016

39:13
and 2017, and that's one of the main justifications for our program.

39:18
The actual official name of our program and our funding agreement is Real Time

39:22
Operations Enhanced Delta Smelt Monitoring Program.

39:25
So it's something that we really, strive to do.

39:29
Just to give you a sense of the scope of the program,

39:32
since we are first established in the summer 2016, we've,

39:35
individually identified almost half a million fish.

39:39
This is a Delta smelt monitoring program, but we capture pretty much

39:42
any fish species in the delta.

39:44
And since we're sampling Delta smelt throughout their lifecycle,

39:47
we use Kodiak trawls to sample juveniles, adult and adults.

39:51
But we also use 20 millimeter trawls to sample larvae.

39:55
And, as you know, larval sampling has a whole other processing component.

39:59
Not just the data collection, not just the field collection

40:02
data entry, but there's also the lab processing as well.

40:05
We also catch half a million.

40:07
We've also caught half a million individual shrimp and jellyfish too.

40:12
So that's a lot of data that we have to report

40:15
and this is probably hard to read,

40:17
but it's just meant to convey the complexity of the data workflow.

40:21
EDSM share staff with the Delta Juvenile Fish

40:23
Monitoring Program, which is just as big of a program as EDSM.

40:27
And so, we have a lot of different things going on.

40:30
Any given day, we could have 20 something field crew in the field

40:35
collecting data, 2 to 4 crew in the in the office

40:39
entering that data for data managers, biologist, supervisor.

40:42
So it's a huge program that takes up a lot of resources.

40:46
And a lot of those resources are, devoted to the whole data process

40:51
and data reporting.

40:55
And so when EDSM was first established,

40:57
we had to kind of think of some new ways to get

41:00
our data entered and reported quicker so we could meet that real time component.

41:04
One thing we decided to do was to take pictures of our data sheets in the field.

41:08
So instead of, you know, writing data on a data sheet, coming back,

41:12
giving it to somebody, they enter it later.

41:14
We would take a picture almost immediately after the sample was collected,

41:18
send it back to the office, and the people in the office

41:20
would enter that data pretty much right after it was collected.

41:24
And we also established this web services get data portal online.

41:29
And so,

41:30
traditionally we would enter our data into an Access database on a server.

41:34
You had to be in the office.

41:35
It was very restrictive.

41:37
And we were able to work with IT for U.S.

41:40
Fish and Wildlife, IT in Denver to create this online portal.

41:42
So you could enter data anywhere from the office or at home.

41:45
But it also created a back end for data users to have access to our database.

41:50
And this has been essential for our partnerships

41:53
with Bay Delta Live and SacPAS.

41:54
So it's a way for them

41:55
to get automatic access to our data and report our data in real time.

41:59
So this was pretty essential to our real time reporting process.

42:04
So in general, how to agency share data.

42:07
Sharing data is a big part of the IEP.

42:09
As we've been talking about.

42:10
the DUWG obviously deals with the whole data lifecycle,

42:14
but then the science synthesis team also deals with with data set integration.

42:18
That's something I'll talk about later.

42:20
And of course, the EDSM isn't the first program to ever worry about sharing data.

42:24
CDFW has a great website with interactive maps and catch

42:28
data, an FTP site where you can download curated databases.

42:33
But how does EDSM specifically share data?

42:36
This is kind of a summary of pretty much everything we do.

42:39
I'm going going to go into most of these into a little more detail.

42:43
The first way that we really established our real time

42:46
data reporting was through what we call our daily report.

42:49
And so this was really manual when EDSM first started before

42:52
we figured out how to make things a little bit faster.

42:57
Data sheets would come back at the end of the day.

42:58
Someone would manually enter this into an Excel spreadsheet

43:02
because it hadn't been entered into a database yet.

43:04
Then they would summarize it, and then they'd send it out to a listserv by email.

43:09
And this is to pretty much anyone who was interested in EDSM data.

43:12
Managers, other agencies, pretty much anyone who wanted to be on it.

43:16
It's also posted every single day to our Fish and Wildlife website.

43:20
You can see the link below.

43:23
And we've worked

43:24
to automate this a little bit over the years to, you know,

43:28
like I mentioned, we take the data sheets and we enter those quicker so that,

43:32
now our effort is centered

43:35
on getting the data into the database and QC'd as much as possible.

43:39
But by the end of the day,

43:40
then we have an R script that we just run

43:42
that takes, you know, just a couple of minutes, generates a file,

43:45
and then we put it into an email and send it out.

43:47
So we've developed ways to make that a lot faster over time

43:50
and kind of reduce that manual component.

43:54
Another one of our major products is our weekly Delta Smelt Abundance report.

43:58
This kind of translates the catch data into actual population

44:02
abundance information.

44:03
And this is really useful to managers who want to know what's going on

44:07
with the Delta smelt population.

44:09
It's intended to tell you exactly how many or, you know, within

44:12
a confidence level, how many smelt are actually out there in any given week.

44:17
and it's also really useful because it has these nice catch maps

44:20
that tell you where we sampled, where delta smelt were caught,

44:23
where they weren't, how abundance indices have changed over time.

44:27
And this report's generated by Laura Mitchell,

44:29
a statistician in our office, and also posted to our website.

44:33
And it's generated every week.

44:37
Another

44:37
curated report that we have is for the Smelt monitoring team.

44:41
We send this out every week during the operate the water operation season.

44:44
And, we really tried to tailor tailor this report for what the smelt monitoring

44:49
team is looking for.

44:50
and it help to help aid them in their recommendations

44:53
to the water operations management team.

44:55
So it has things like the total delta smelt catch, total catch

45:00
total over the water year, catch maps with the delta small and longfin smelt.

45:05
just anything that we think would be useful to the smelt monitoring team.

45:08
And this goes out every week.

45:13
We also post our data in pretty much real time on Bay Delta Live.

45:17
Like I mentioned before.

45:18
this has been this is through the web services

45:21
direct link to our online database portal.

45:24
and this is really great because it's all automated.

45:28
The data is reported almost immediately, and we've been going through an effort

45:33
lately working with Bay Delta Live to curate our page

45:36
and kind of make it mimic some of the other reports

45:39
that we have, like our daily report and even our smelt monitoring team report.

45:44
So we can again reduce the manual component and spend

45:48
a lot of the effort upfront, designing the tables, deciding what we want to show,

45:53
making it completely automatic, automated and then reducing that manual component.

45:57
So that's kind of where we're heading.

45:59
so maybe instead of having to run a report every single day for the daily report,

46:04
we could just have it automated on Bay Delta Live and send people there.

46:08
That'll make our data more accessible, because not just people who know about it

46:11
and are on the listserv, but it's pretty much anyone

46:14
who goes to Bay Delta Live or happens upon it on the internet.

46:17
And also, people who are looking for it.

46:19
So this is going to make our data a lot more accessible,

46:22
also a lot more automated. So easier for us.

46:26
Bay Delta

46:27
Live is great because, it has all the tables, it has all the

46:31
all the catch information and also has these really nice interactive maps.

46:35
Catch data, not just delta smelt,

46:37
but pretty much any species that you want to look at.

46:41
Similarly, our data is also posted on SacPAS

46:44
for some of their, salmon and smelt monitoring team pages.

46:48
And they have curated tables and they also have access to that web services link.

46:52
So direct access to our database.

46:55
And it's automated automatically in real time.

46:59
And everything I've talked about before is real time preliminary data.

47:02
But we also post our data on, EDI, which we'll talk which,

47:07
you know, we'll be talking

47:07
about later in the presentation, but this is where you want to go.

47:10
If you actually want to use EDSM for a scientific product

47:13
that you want to publish,

47:14
this is gone through the whole QQ process, but it's not in real time.

47:19
It takes almost a year to get up there.

47:20
but this is going to be the most accurate data.

47:24
Everything else, you know, hasn't really gone through the QC process,

47:27
but, it's out in real time.

47:31
And one other thing that

47:33
EDSM does to share data is integrated data sets.

47:37
EDSM

47:38
is part of a lot of collaborative groups that are working on delta smelt research.

47:43
One is the smelt supplementation effort.

47:46
And we've helped curate a data set that ties together

47:49
all delta smelt catch data since, supplementation

47:53
began in December 2021.

47:56
And this ties together any delta

47:58
smelt catch data from any agency, salvage,

48:01
pretty much anyone seeing Delta smelt and we put it all

48:05
in the same format and into one data set.

48:08
And Rosie Hartman has turned that into a cool, shiny app where you can also see

48:12
some interactive maps and, look at the data on a map.

48:17
And we're also part of the Directed Outflow Project,

48:19
which is a collaborative group that does a lot of delta smelt research.

48:23
And this this diagram is meant to convey the complexity of everything

48:26
that happens to every single delta smelt caught by the monitoring agencies.

48:30
A lot of different things happen to them.

48:32
And, the effort here is to try and tie

48:36
the catch data to some of that downstream data.

48:39
This is especially important for something like genetics,

48:42
where that could actually change

48:44
the species ID or change your original catch data,

48:47
and we need a better way to tie them together.

48:50
So we created this.

48:52
We're in the process of creating an integrated, integrated data set

48:55
that will tie together all these different all the catch data to all these

48:59
different downstream data, and then help us get that downstream data

49:03
tied better with our data in our database.

49:08
some lessons learned and where we're going in the future.

49:11
Real time reporting is really useful to managers,

49:14
but it's also extremely resource intensive.

49:17
So, one of the reasons that EDSM requires so many resources,

49:21
but the more efficient you are,

49:22
the more data you'll be able to collect and a higher sample size.

49:26
Usually means higher certainty in whatever you're trying to say.

49:29
Ways to get to that efficiency is automate wherever you can.

49:33
I've mentioned that a few times already, but, you know,

49:35
things like Bay Delta Live, R scripts instead of manual queries.

49:39
Just automate as much as possible and spend the time upfront,

49:43
designing your automation so that it's easier later.

49:47
We use reports wherever possible.

49:49
We do have a lot of targeted reports, but where we can we do try and reuse them.

49:54
So we don't have,

49:55
you know, one specific script or product for every single different thing we do.

49:58
but something that we're looking to do in the future is electronic data entry.

50:04
Of course, electronic data entry has pros and cons to it, but

50:08
when one of your main objectives is real time data entry,

50:11
I think the pros probably outweigh the cons because it takes out removes

50:15
a lot of the time consuming steps, and it's going to help you

50:19
report your data faster.

50:20
And and one thing I think we could maybe improve on

50:23
as a community is integrating data sets earlier.

50:27
I think there's a lot of great work being done to integrate data

50:30
sets after data is collected, after it's gone

50:33
through the QA/QC process. A lot of synthesis efforts.

50:37
But if we tried to integrate these data sets earlier,

50:40
I think it would help, tie data together in earlier stages and help managers

50:45
make decision, getting decisions by having the whole picture

50:49
and everything tied together at a earlier stage of the data collection process.

50:55
And that's about it.

50:56
thank you.

50:57
Thanks, Denise.

50:59
Next up, we have Jay Bosworth from DWR

51:02
talking about coded data access.

51:06
Alright.

51:10
Thanks, Cat.

51:12
All right, so, morning, everybody.

51:14
Dave Bosworth, I'm with the Department of Water Resources,

51:20
and I'm also one of the co-chairs of the Data Utilization Working Group, or

51:25
DUWG.

51:26
So today I'm going to

51:29
we're going to kind of switch gears a little bit and,

51:32
talk about, some get into some R coding here.

51:36
but Denise's, talk was a good segue into coded data access.

51:41
So talking about accessing data,

51:44
using the R programing language.

51:47
So just want to mention that this is a large topic.

51:51
So it's, meant to, this is meant to be like a brief overview,

51:55
kind of beginner to intermediate level.

51:57
And I think that someone might put the link

52:00
to the code that I'll be using in this tutorial in the chat.

52:03
So, you could, open that up and follow along if you would like.

52:09
All right.

52:11
Great.

52:13
So first of all, what do we even mean

52:16
if we're talking about coded data access?

52:18
So, I like to think of it as,

52:21
using code, to import data into R, and it's from

52:25
the key is from an external or non-local and open source.

52:29
This is not importing data from your

52:31
your hard drive or from a shared network or even SharePoint.

52:35
Typically this is actually accessing data open data from an online source,

52:40
such as like an online data portal or data repository.

52:47
All right.

52:48
Great.

52:51
So just to give you a brief overview of kind of

52:54
what we're going to talk about here for the next 30 or so minutes,

52:57
we're going to get into a few simple methods in order to accomplish this.

53:01
then we're going to get into a few

53:03
dedicated R packages that help with accessing data.

53:06
and these are kind of specific to the Bay Delta and IEP.

53:09
And then we're going to get real specific and talk about some IEP integrated

53:13
data sets and some R packages to access those.

53:20
Before we get

53:21
into the code here, I just want to just briefly

53:25
explain what a why even bother trying to trying to do this.

53:28
I find, that this is, pretty convenient way and efficient way of writing code,

53:34
and it's my preferred method if you want to have a reproducible workflow

53:37
since you're accessing data, that's open and available online.

53:43
So another, nice thing is that you're able

53:46
to better keep track of your updates because it's written in your code.

53:50
And so versioning is really nice.

53:52
But, you do want to be aware of changing versions in your, your data source.

53:56
And we'll get more into that

53:57
at the, in the Q&A and kind of how you could deal with, with that kind of thing.

54:04
All right.

54:05
So first we'll talk about a couple simple methods.

54:08
first one is using a read

54:12
CSV function, with a URL as the, the file path.

54:16
And the second one is a two-step method that's first used download

54:21
for the download file function to download your file,

54:23
and then you import it into your workspace.

54:32
All right.

54:34
So let's get into the first one here.

54:36
Using read CSV with a URL.

54:40
So you could use read underscore CSV or read dot CSV, whatever you prefer.

54:45
I like to use the Tidyverse versions in my code.

54:47
So that's kind of what I'm doing here.

54:52
So, if you take a look at this, example here.

54:57
So we have the day flow.

54:58
We have the day flow results for 2023 on the CNRA data portal.

55:03
If you look at this URL to the, to this, that links to this data,

55:08
you'll see that, it points to a CSV file.

55:10
So, this is really a great workflow when your, your URL points to a CSV file.

55:18
So if we copy this, URL.

55:24
And then we paste it into, the read csv

55:30
function here as the path.

55:34
If you get your data

55:35
imported, into R and you can start working with it.

55:38
So as simple as that.

55:42
Let's see I realize it'd be nice to have a laser pointer here.

55:46
There you go.

55:48
Real cool.

55:49
All right. So the second simple method here

55:52
is download dot file and then import.

55:56
So this this may

56:00
this might be necessary this might be a necessary workflow.

56:03
when you, are working with certain type of file types

56:07
from online sources like, excel files, PDF or ZIP files, it might be

56:12
it might be required to actually download the file first

56:16
to your, local computer and then import it from there into R so,

56:20
we'll kind of go with another example here.

56:24
And we're going to work with the environmental monitoring

56:27
program, discreet water quality, data package.

56:30
This is on EDI or environmental, the environmental data

56:33
initiative we've been talking about before.

56:36
And so if you go to this webpage and you right click on the download

56:41
button here for this first data file,

56:45
and then you click on copy link address.

56:48
Then you get, this nice, URL here.

56:54
So then if you actually if you paste that in or you use that,

56:57
URL in this download file as your path,

57:01
and then you the second argument here in this download file,

57:05
function is your where you're saving this file to temporarily.

57:10
So I normally use a temporary directory so it doesn't stick around for too long.

57:13
And then you just give it a name whatever you want.

57:16
So once you do that, it downloads the file for you.

57:19
And then the second step is,

57:22
using a read CSV function, and giving it

57:26
the path to that temporary file that you just downloaded.

57:30
And then voila there you go, there's your data.

57:35
All right.

57:37
So, next up, I'm going to cover,

57:40
three dedicated R packages, that help you access,

57:46
online data.

57:47
And, the first one is going to be we're going to cover is edi utils,

57:52
and that's for, kind of working with the Environmental Data Initiative.

57:56
Second one is data retrieval, and that's working with USGS,

57:59
NWIS data or the National Water Information System.

58:03
And then the third is Cedar.

58:05
that's helps you access, the data from CDEC,

58:09
California, the California Data Exchange Center.

58:13
So first up is edi utils.

58:15
like I said, this is an API, client

58:18
for the Environmental Data Initiative, or EDI, that we call like a call here.

58:22
so this allows you to search, access and upload data,

58:25
from EDI into your R environment.

58:29
These slides will be made available after the workshop.

58:32
So I wanted to include some links to some documentation

58:35
and installation references here for, for future.

58:39
So you could you could click on these later.

58:44
All right.

58:44
So I'm going to highlight

58:46
some of the functions in this package by running through an example.

58:49
And the example I'm going to use is again the discrete water quality

58:54
data collected by the environmental monitoring program,

58:58
at IEP.

59:01
And this data is found on that,

59:05
as a, as a data package on EDI.

59:07
So if you go to this, URL and we do a simple search

59:11
and we search for IEP and EMP, we get a table of results.

59:15
And if you take a look at the third result down, this is the data package

59:19
we're interested in.

59:21
And you off, over to the right you can see there's a package ID that's

59:24
EDI 458 .10.

59:30
So just got to remember that, so,

59:33
so to start this example off, you know, will load the package and,

59:38
define a few, elements that will help us, in the later, later in the example.

59:44
So EDI scope is EDI, and then the package ID

59:47
458.

59:56
All right.

59:58
So we have that we have our, our you find R package ID and all that.

1:00:03
And now we kind of want to know

1:00:05
what is what are the all the revisions that this package has gone through.

1:00:08
So we could use this list data package revisions function.

1:00:14
If we use that and we define the scope and the identifier

1:00:17
that we just defined earlier, we just defined earlier,

1:00:21
it'll give you a vector of of package revisions.

1:00:25
You can see there has been ten package revisions.

1:00:29
And if you actually went to the website you would see that

1:00:31
this this is confirmed here.

1:00:33
This is what you would see.

1:00:37
So a nice optional argument

1:00:40
for this data package revisions

1:00:42
function is this filter equals newest argument.

1:00:46
that actually just gives you the latest the latest version number.

1:00:50
So in this case if we return the number ten

1:00:53
and then we can just kind of use a little, paste and paste these elements

1:00:57
together to create an EDI EMP package ID.

1:01:02
So we'll do that and we'll define that

1:01:06
for later.

1:01:09
Next up is we want to know okay.

1:01:12
What are what are the some of the data components within this data package.

1:01:15
There could be more than one, data component or data entity

1:01:19
within a data package.

1:01:20
So we kind of want to see what would exist here in this, this data package.

1:01:24
So here's our code from before.

1:01:26
And if we use this re data entity names function.

1:01:30
And we provided the package ID that we just created,

1:01:34
it'll provide this, this data frame for you

1:01:37
of all the data entities that exist in this data package.

1:01:40
So it's a, it's a two column, data frame.

1:01:43
The first is STID that kind of provides this long hash code for that,

1:01:47
that helps, link or provide a link to the, the, the data.

1:01:53
and then this,

1:01:54
this entity name is actually kind of what you want to look, you looking for.

1:01:57
This is actually kind of a more user friendly description of, of the, of

1:02:00
the data of the data entity.

1:02:03
If you were to go to the website again,

1:02:06
this is what you would see.

1:02:07
And that matches what we have here.

1:02:09
So just kind of proof of concept here.

1:02:11
Just making sure that it's doing the right thing.

1:02:15
So okay, continuing on with this, example.

1:02:19
So now we want to read, some data into our R environment

1:02:23
so that we could use the read data entity, function and edi utils.

1:02:28
we really were interested in this first, first entity

1:02:32
here, the EMP, this, discreet water quality data.

1:02:36
so we were gonna focus on that one.

1:02:39
If we just pull that, entity idea

1:02:42
ID just for that first, that first element,

1:02:46
and then we, use the read data entity,

1:02:48
function and provide the package ID and the entity ID

1:02:53
within that function, it'll, it'll, import the data.

1:02:57
But, notice this is a two-step process because

1:03:00
read data entity imports your data as raw bytes.

1:03:03
And so you need to use a reader function with it.

1:03:05
In this case we're going to use read CSV.

1:03:07
So we we do that.

1:03:08
And then there you go. There's your data.

1:03:13
All right

1:03:14
so kind of a quick overview of EDI utils.

1:03:17
there's a lot more in there.

1:03:18
So check it out if you're if you're curious.

1:03:21
So the next R package is data retrieval.

1:03:25
And this is, allows you to, to load USGS data into your R environment.

1:03:29
And it has both, NWIS and WQP web services.

1:03:33
And this is the National Water Information system and WQP stands for the Water

1:03:37
Quality Portal

1:03:38
and that has data from USEPA, USDA and USGS.

1:03:43
And again, here's some package documentation for you.

1:03:48
So we'll start off,

1:03:51
maybe focusing on the NWIS web services side of things.

1:03:54
and I'll very briefly show you how to use a few of the WQP web

1:03:58
service functions later.

1:04:00
So again, we're going to,

1:04:02
walk through some of the package functions by running through an example.

1:04:06
and so we're going to work with this Sacramento River at Freeport USGS station.

1:04:11
And just something to note here.

1:04:13
each USGS station station has a unique, eight digit, identifier.

1:04:18
And that's this is the identifier for the station that we're interested in.

1:04:21
And for the purpose of this example, we're going to we're going to want to

1:04:24
we're going to work with, instantaneous or 15 minute,

1:04:28
specific conductance data collected there.

1:04:31
So first thing you want to know is maybe like some site information

1:04:34
about your, your, your, your, your, station.

1:04:39
So, you could use the whatNWIS sites function,

1:04:43
to get some information when you, when you run that, it'll provide,

1:04:48
a handy little table for you and provide you some lat longs

1:04:54
and, some other metadata for that station.

1:05:00
Next thing you might be wanting to know

1:05:02
about your your, the station is what data is available there.

1:05:06
And so the whatNWIS data function allows you to,

1:05:10
to gather some information about what data is being collected.

1:05:14
And so, we'll use this

1:05:17
what NWIS data, function and you provide the site number

1:05:21
and then this is an optional, argument for service.

1:05:25
And that stands for UVs stands for Instantaneous data.

1:05:28
So we're just focusing on the instantaneous data or 15 minute data

1:05:31
that's collected here.

1:05:34
and kind of puts off a pretty large table.

1:05:37
So I use the Glympse function to kind of show you, show you all of it.

1:05:40
But I just want to focus, highlight a few things that are

1:05:43
there are a few elements on this table that are, that are useful.

1:05:46
First is the parm CD column or that stands for USGS Parameter code.

1:05:52
You can see that there five digit codes.

1:05:54
and unless you really,

1:05:56
you know, are familiar with USGS data, those don't really mean much to you.

1:05:59
But we'll get into that later to how to decode those.

1:06:03
The second thing I want to point out is,

1:06:06
at the very end, we have, begin date and a and count number.

1:06:10
So this is kind of the period of record.

1:06:12
The sample count for each of these parameters at the station.

1:06:15
So it's kind of nice to know.

1:06:17
So now we're going to get into these parameter codes.

1:06:21
So within the data retrieval R package, there

1:06:25
there is a file or a table

1:06:28
called parameter cd file.

1:06:31
and you just print that off in R console,

1:06:35
it provides a complete list of the USGS parameter codes.

1:06:38
And you can see that there is a column for the each parameter code.

1:06:42
And then kind of like a nice description of actually what it stands for

1:06:46
and the units that it's then. So this is pretty helpful.

1:06:50
but I guess, you know, it's

1:06:51
not as helpful as you want, maybe because it has all of the parameter codes.

1:06:55
So, you know, you could you could search this table

1:06:58
to learn more about the parameter codes that you're interested in.

1:07:02
One way of doing that is by using this link here

1:07:05
that I provided to a USGS website.

1:07:08
and this helps with searching for parameter codes.

1:07:11
you could also just use some R code to filter your this table down.

1:07:15
And this is the this is the filtered, code list.

1:07:20
and this is for, just the, the parameters that are collected at the, the,

1:07:25
the Freeport station that we're interested in.

1:07:29
So like I said, we are interested in specific conductance.

1:07:31
So I highlighted that here.

1:07:34
Our parameter code that we need to use for that is 00095.

1:07:40
So now we're ready

1:07:41
to, read our, our data in, to R workspace.

1:07:44
So now we're going to use the read NWIS UV function.

1:07:47
the UV I guess stands for instantaneous data.

1:07:50
Or maybe it's like unit value.

1:07:52
I'm not sure, but, this is the function you'd use

1:07:55
to import instantaneous data for, for one or more stations.

1:07:59
And parameters from NWIS.

1:08:00
In our example here, if we were providing our site number,

1:08:04
our parameter code, and then there are some optional arguments here

1:08:09
that you can specify your start and end date and your time zone.

1:08:12
If you just left those blank, it basically give you the entire period of record.

1:08:16
And the UTC time zone.

1:08:21
You run this,

1:08:22
it does the import for you and puts it into your R workspace.

1:08:26
and here's your data.

1:08:28
so this column here starts with an X and has the, the parameter code.

1:08:32
This is, this is your, your your values, your data values.

1:08:36
And then I just wanted to note that that just to the right of it this there's

1:08:39
this column that ends with CD and this is the quality code, for this data.

1:08:45
And so typically there's a, there's like

1:08:47
just a couple of quality codes in this and USGS data, one is approved

1:08:51
and that's A and P is stands for provisional.

1:08:55
and you can see this is provisional.

1:08:57
Just something to keep in mind.

1:08:58
Sometimes provisional data can can is subject to change.

1:09:01
I did just want to note that that you could use more than one station

1:09:05
and parameter in this function.

1:09:06
So if you wanted to do that, you could just provide a character vector

1:09:09
for each one of these arguments with more than one code.

1:09:13
That's how you accomplish that.

1:09:15
I'm not going to get into it too much, but, I just wanted there's

1:09:18
kind of like a related function to, this instantaneous data function.

1:09:23
It's for importing daily data.

1:09:24
And this is, it's read NWIS is DV,

1:09:27
and so it basically has it works the same way.

1:09:30
but there's an additional, stat CD argument and that,

1:09:35
that defines the, the, the aggregation

1:09:38
function, for, for the daily values.

1:09:42
And so if you don't provide anything for

1:09:43
that, it's just going to give you daily means.

1:09:45
But if you specifically wanted something else you can provide that code.

1:09:49
And here are some of the common ones you might want to use.

1:09:52
And here is a link to all of them

1:09:54
if you're so curious.

1:09:58
So now we're just going to briefly

1:09:59
just whiz by a few of the WQP functions.

1:10:03
because I don't have a ton of time here.

1:10:06
So, first, the first one I want to know is if you want to

1:10:11
just get station information, you could use the whatWQP sites.

1:10:15
It's pretty similar to what I just showed you before for the NWIS

1:10:19
side of things.

1:10:22
If you want a data availability table,

1:10:24
you could use this three WQP summary function.

1:10:30
And if you want to import,

1:10:32
data, you would use this read

1:10:36
WQPqw function.

1:10:38
They all work pretty similar to the other functions I just showed you.

1:10:42
I did want to note that all for all the WQP functions,

1:10:45
you need this USGS dash prefix for the site codes.

1:10:49
So you can see that here.

1:10:50
I provided that in all of these.

1:10:52
That's that's one minor difference.

1:10:55
that's about all I got for, for for the WQP functions, if you want.

1:11:00
if you want to learn more, obviously refer to the package documentation

1:11:04
for more information.

1:11:06
And I just also want to note that for if you're,

1:11:09
if you're interested in downloading, discrete water quality data

1:11:12
that USGS collects, this is the WQP function

1:11:17
is what you're going to want to use for that.

1:11:22
All right.

1:11:23
Now we're going to get into the last dedicated hour package at Cder.

1:11:27
This is a, a web IPA,

1:11:31
API client for the California Data Exchange Center, or CDEC.

1:11:35
Again, this allows you to import data into your environment from CDEC.

1:11:38
And here's your package documentation and installation, information.

1:11:44
I do want to note that there are a few other R packages, out there

1:11:48
in the wild that allow you to interact and download access data from CDEC.

1:11:52
one of them is CDEC

1:11:53
Retrieve and another one is sharpshoot R.

1:11:56
But I'm more familiar with this one.

1:11:57
So this is what we're going to cover in this tutorial.

1:12:02
So first thing you might want to know

1:12:04
is kind of some basic information about your station from CDEC.

1:12:08
and so again we're going to walk through just a real brief example.

1:12:12
In this case we're going to work with the Sacramento River at Freeport Station.

1:12:15
Again I hope you just worked with with the prior example.

1:12:18
the, the three, letter code, for that station in CDEC is FPT.

1:12:24
So if we provide that to the CDEC meta function,

1:12:28
it'll actually open up a web browser window for you.

1:12:31
And it will provide, kind of basically a it's the web

1:12:35
browser window that has all the station information that you might want to use.

1:12:39
it has some location info and then it has some data availability info.

1:12:43
What's collected there, what's the sensor number and duration and all that.

1:12:48
It's kind of nice to know this, for when you're setting up your data query.

1:12:52
So for this example, we're going to work with, turbidity data

1:12:56
collected here at the station.

1:12:57
And you can just keep a note that the sensor number is 221.

1:13:01
And we're going to work with event data.

1:13:06
So next up is downloading this data from CDEC.

1:13:09
So you could use the CDEC query function.

1:13:12
and so when we run this

1:13:15
we'll, we'll specify our station here at with the station code FPT.

1:13:20
And then this is our sensor number that we just, I just pointed out 221.

1:13:23
And then our durations is a e standing for event.

1:13:28
if you wanted a different, duration

1:13:30
value, you could provide it in this argument.

1:13:33
here is some, here is the ones that are available.

1:13:37
You could also have the option to provide start and end date again

1:13:41
if you, if you don't provide those, I think it would end up just giving you

1:13:45
the entire period of record, which is probably would take

1:13:48
a really long time because there's a lot of data on CDEC.

1:13:52
again, like the USGS, package, you could,

1:13:56
you could use this,

1:13:57
function for more than one station and sensor at a time.

1:14:01
So in that case, again, you would just provide a character vector

1:14:04
with more than one element for the stations and sensors arguments here.

1:14:10
When you run this here, it provides your,

1:14:13
your data and, and you're ready to get ready to go.

1:14:18
All right.

1:14:18
So, last up, we're going to cover some, of the IEP integrated data sets.

1:14:24
and some R packages that allow you to, to work with them.

1:14:28
So we're going to talk about these four here.

1:14:33
I just want to mention that, these are not available on Cran.

1:14:36
So you need to use the dev tools install GitHub function to install these packages.

1:14:43
And each each GitHub

1:14:44
site, for these packages has instructions on how to do so.

1:14:48
and additionally, if you have a windows computer, like many of us do,

1:14:52
you also need to have R tools installed, which is not R it's

1:14:56
actually a separate software.

1:14:58
and you need to have this installed.

1:14:59
And so in order to do that, I just you could visit this, link

1:15:02
and it gives you some instructions and how to download and all that.

1:15:08
So we're going to get to the first one here, deltafish.

1:15:10
So this provides an easy query access to the very large

1:15:16
publish EDI data set of IEP fish abundance and length data.

1:15:20
this is a long term data set that has data from nine surveys.

1:15:24
And here's alphabet soup here for you.

1:15:26
I'm not going to read these,

1:15:27
but these are all the surveys that are available, in this data package.

1:15:31
and then here is the link to the this R package on GitHub.

1:15:35
And then this is the link to the actual data repository.

1:15:38
It's pulling data from.

1:15:40
This is a really nice package if you're just wanting a subset of this data.

1:15:44
If you were to download this data and try to work with it in R,

1:15:47
and it takes an extremely long time.

1:15:48
So this is actually like a, like a time saver,

1:15:51
kind of helps you do things a lot more quickly, access

1:15:54
and query data a lot more quickly.

1:15:57
So first step for working with this

1:16:00
data package is we're going to build and load the database.

1:16:04
And so,

1:16:04
when you load deltafish and then you you run this function called create-fish-

1:16:10
db this will actually cache a local copy on your computer.

1:16:15
And this is actually this is actually, only necessary to do

1:16:18
once for each EDI update, after you install this package.

1:16:23
so it's kind of nice because it does take quite a while.

1:16:25
If you do know that there is a new version, on EDI,

1:16:30
you could use the update equals true argument within this function,

1:16:33
and then it will it will rebuild the cache database for you.

1:16:37
Next up is opening just the two data files that exist in this database.

1:16:41
And so that's using the functions open survey and open fish.

1:16:45
And we'll just give we'll name them surv and fish here in our R environment

1:16:48
so we can work with them.

1:16:52
Then we

1:16:52
want to build and run our query from this database.

1:16:57
So the nice thing about about this,

1:17:00
this are packages that you could use, dplyr functions to build your query.

1:17:04
So if you're, you know, fluent in the dplyr r package

1:17:08
from the tidyverse, then this is pretty simple to do.

1:17:11
And you don't have to know actual SQL programing language to, to do this.

1:17:15
so in this example we're going to we're going to pull some data

1:17:19
from the fall Mid-water trawl, and from the survey table.

1:17:24
And then we're going to, filter to just three fish species from the fish table.

1:17:30
And then we're going to join them together,

1:17:32
using the a left join, which is also from dplyr.

1:17:37
and that basically kind of builds your query.

1:17:41
but if you actually want to run the query and import the data

1:17:45
into R you would use this dplyr collect function.

1:17:49
once you do that, it runs the query, imports it in.

1:17:53
And here's your data.

1:17:55
but I it's probably much faster than doing this from,

1:17:58
from the entire data set.

1:18:01
All right.

1:18:02
Next up, we're going to talk briefly about the zooper R package.

1:18:06
Tis allows for, you to download and integrate, the IEP.

1:18:12
zooplankton data from seven surveys and that again, a very long

1:18:15
term, record here, again available on GitHub.

1:18:19
And then I just want to note that this data set is periodically published on EDI.

1:18:25
And here's a link to that, as a static version.

1:18:28
If you want the most recent zooplankton data that you would want

1:18:32
to use this R package to do the downloading integration for you.

1:18:36
Most of us are just

1:18:38
going to use one function from this package its Zoopsynther.

1:18:41
This does the downloading and integrating for you.

1:18:43
but a couple of I just want to point out a couple of the,

1:18:46
arguments that you could specify within this function that are important to know.

1:18:51
First is data type

1:18:52
and you could use

1:18:54
you need to use this argument to choose between the two approaches that this

1:18:57
R package uses to resolve differences and taxonomic resolution.

1:19:02
So if you use the taxa to argument here

1:19:05
for data type, it will give you all available data on the given taxa.

1:19:09
But if you want to do more of a community analysis,

1:19:12
then you should use the community.

1:19:13
argument here.

1:19:15
I also want to mention a few other things you can specify,

1:19:19
such as response, the response variable.

1:19:21
in this case for you could we're doing the catch

1:19:24
per unit effort and biomass per unit effort effort.

1:19:28
So you can you could specify that also, the sources that you want to integrate

1:19:33
and the size classes that you, that you're interested in and a date range

1:19:38
when you run that does the downloading and integration for you.

1:19:41
And and then you're ready to, to move on to your analysis.

1:19:48
So another R package that helps you, work

1:19:52
with IEP integrated data sets is the discrete WQ r package.

1:19:56
this is for, water quality data or discrete

1:20:00
water quality data that, all the IEP agencies collect.

1:20:04
I, um, this uh has a ton of different surveys in it,

1:20:08
for a very long period of time.

1:20:10
Again, gigantic alphabet soup here.

1:20:12
But if you're interested, there they are. There are

1:20:14
All the surveys

1:20:14
that are in this, this integrated data set, it's available on GitHub.

1:20:19
And then again, like the zooper package, this data set is periodically published

1:20:23
on EDI as a static version.

1:20:24
But if you want

1:20:25
the most recent version of this data, you should just use this R package to do

1:20:29
to get your data.

1:20:32
The function that you would use

1:20:34
to do your access your data access is the WQ function.

1:20:37
again, you can specify the data sources that you're interested in and your start

1:20:41
and your end here. Pretty simple.

1:20:44
when you do so it'll, it'll import your data

1:20:47
for you that you, you queried.

1:20:51
Last integrated,

1:20:53
IEP data set that I want to touch on is, the deltamapr.

1:20:59
R package, this provides, spatial data for the SF Bay

1:21:03
Delta this is

1:21:04
this is something I use a lot for maps for doing spatial analysis.

1:21:08
All the data objects in this R package are stored in the SF format,

1:21:12
or that stands for Simple Features, or I like to think of it as spatial features.

1:21:16
and there is, for as four kinds of main data types.

1:21:21
So we have waterway data, region data, habitat data and station and point data.

1:21:27
And each of these are have a prefix to kind of they kind of denote which,

1:21:31
which kind of data type it is.

1:21:33
if you're curious about this package and what data exists in it

1:21:37
that I definitely recommend to, to use this link and check it out,

1:21:40
there's a helpful Readme file.

1:21:45
just to kind of show you

1:21:47
an example of one of the data objects that are within this package.

1:21:50
here is the wwdelta

1:21:54
data, spatial data, object.

1:21:57
This is the Delta waterways.

1:22:02
When you load this into your R console

1:22:05
or your R session, it kind of, you know, it provides a table here, right?

1:22:09
So it's that's, I guess, okay.

1:22:11
But, really, the beauty of this is that you could then use these,

1:22:16
data objects to create maps or do spatial analysis.

1:22:19
So here is just a real quick example of creating a map of this,

1:22:23
of this spatial data.

1:22:24
And you can see it's a nice space map of, waterways and the, the bay.

1:22:29
And you could use this for, for other for, for your maps.

1:22:32
This is great.

1:22:36
So that's about it.

1:22:37
There is a, there's a lot of other, topics

1:22:41
we could get into here, but, I just kind of wanted to give you a brief overview.

1:22:45
So a few other resources here

1:22:48
that might help you in your, coded data access,

1:22:51
journey, as if you were interested in webs and web scraping.

1:22:55
I highly recommend checking out the rvest package.

1:22:57
If you have a PDF file that you have online

1:23:00
that you want to extract data from, then you.

1:23:02
I would suggest trying to use the PDF tools package.

1:23:06
The content ID package is kind of an interesting

1:23:10
one that helps you work with external data through content identifiers.

1:23:14
and we'll get that into that here.

1:23:15
But it's it's kind of

1:23:16
if you're curious about that, that didn't check out this link.

1:23:19
And then here's the code from this tutorial.

1:23:23
And again this is on GitHub.

1:23:24
So if you want to go back to any of that code that I, that I showed,

1:23:28
this is where you could find it.

1:23:29
And that concludes my my talk.

1:23:31
I hope you found it to be useful.

1:23:40
Awesome.

1:23:40
Thank you Dave and thanks, Rosie for posting all the links.

1:23:44
Yeah, thanks. That's a lot of links.

1:23:45
A lot of links.

1:23:47
Yeah, yeah.

1:23:48
We will, like I mentioned, send out the resource spreadsheet.

1:23:54
It might have been posted in this, chat already.

1:23:57
of, of a lot of these links.

1:23:59
So you'll be able to refer to those later, but feel free to ask if,

1:24:02
if you need a link repasted.

1:24:06
So We're going to move into the next session of today's workshop,

1:24:10
which will be focused on metadata and data publication.

1:24:14
So I'll turn it over to Trinh to share his presentation.

1:24:22
All right.

1:24:24
So welcome, everyone.

1:24:25
this is our data legacy

1:24:28
metadata and data publication best practices.

1:24:31
My name is Trinh Nguyen.

1:24:32
I'm a senior ES specialist working here at the Interagency

1:24:36
Ecological Program within the California Department of Fish and Wildlife.

1:24:41
And by the end of this training.

1:24:46
I hope to be able to provide you all with a detailed

1:24:48
look at what metadata and data publication best practices are.

1:24:53
I'll provide you

1:24:54
with some practical guidance to these concepts on how to apply them,

1:24:57
and also highlight some opportunities to work with the IEP

1:25:03
on, incorporating these concepts to our system.

1:25:09
That being said, though, I'd like to just note

1:25:11
that this is an abbreviated version of a longer training that

1:25:17
I would like to put on later on this year.

1:25:19
And in that training, I will go more in depth

1:25:23
on the concepts that I introduced today.

1:25:26
There will also be a workshop component in which, I'll have more time

1:25:31
to showcase some of the practical guidance that I highlight today.

1:25:35
And also, you'll be able to bring your own data

1:25:38
sets to apply these concepts to them.

1:25:42
And so if that sounds useful, please keep an eye out on your inbox

1:25:46
and I will send out that email when I am ready.

1:25:51
For today's training

1:25:53
though, I'll first go over the vision that I have for why

1:25:57
I think these concepts are so important for our system.

1:26:03
I'll then go into defining what metadata

1:26:05
and data publication best practices are.

1:26:09
What are their benefits?

1:26:10
How do we accomplish them?

1:26:13
And then I'll end talking about the culture that we have to develop

1:26:17
and embrace.

1:26:18
If we want to be successful in incorporating these concepts.

1:26:24
And so, as many of you all know, IEP collects

1:26:27
a lot of data and we collect a lot of data across space and time.

1:26:32
Meaning we collect a lot of data at different regions within the estuary

1:26:36
throughout the estuary.

1:26:38
And we also collect data

1:26:39
during different times of year and throughout the entire year.

1:26:43
And so this results in us having collected a lot of raw data every season.

1:26:51
But having a lot of raw data doesn't necessarily

1:26:54
make it automatically into useful data.

1:26:58
Data.

1:26:58
is only valuable and impactful if if used

1:27:03
and if it's used correctly.

1:27:07
And so here I have a diagram of three levels

1:27:09
of data raw data, usable data and useful data.

1:27:13
And to go from one to the next, there are various requirements.

1:27:17
So to make your raw data into usable data,

1:27:21
your users have to be able to discover it.

1:27:23
They have to be able to access it,

1:27:25
and it has to be relevant to the questions that they have

1:27:29
to go from usable data to useful data.

1:27:32
You have to have documentations to

1:27:35
guide your users on how you can use your data correctly,

1:27:38
how they can interpret your data correctly to get to the correct conclusions.

1:27:43
You also have to update your data regularly so that your users can answer

1:27:47
relevant, up to date questions that they might have.

1:27:52
And so obviously, there's a lot of extra work

1:27:54
that you have to do beyond just collecting and publishing the data

1:27:58
if you want to make your raw data as usable and useful as possible.

1:28:05
And that's one of the main difference between

1:28:08
what a data producer does and what a data steward does.

1:28:13
And so for a data producer, they go out and they collect a data

1:28:16
and they work up and QA/QC the data, and then they publish that data.

1:28:21
But how the users find that data, access that data, use

1:28:25
the data is beyond the scope of what a data producer does.

1:28:31
A data steward, on the other hand, is a data producer,

1:28:35
but they continue to be a conduit

1:28:38
between their data users and their data.

1:28:42
Essentially, they can do this directly, in which the users come to them

1:28:46
to ask questions about how to find, access and use their data,

1:28:50
or they can do this indirectly, in which documentations are created

1:28:55
to be published alongside the data set that then guides their users

1:28:59
on how to use their data set correctly.

1:29:02
And I think this shift in viewing ourselves

1:29:06
not as data producers, but as data stewards,

1:29:09
is one of the most important thing that we need to do

1:29:12
if we want to make our data as usable as possible.

1:29:18
And ultimately,

1:29:19
I think this is something that is really important for us to do.

1:29:23
You probably heard, the phrase best available science

1:29:26
and how our management actions should be based on best available science.

1:29:31
And that makes sense, right?

1:29:33
Because it's probably then that's theoretically and the management

1:29:37
action will be the most effective when we best understand our system.

1:29:42
What you may not have heard of as often is, though,

1:29:45
that the best available science actually depends on the best

1:29:49
available data, and so ultimately

1:29:52
it all begins with the raw data that we collect,

1:29:56
and it is through this raw data that that shapes

1:29:59
our understanding of our system.

1:30:01
It is through that understanding

1:30:02
of our system, then, that shapes how we can manage our system.

1:30:06
And if we manage our system correctly.

1:30:11
The other point, though, of why I think it's so important to put in this

1:30:14
extra work is due to this process

1:30:17
called information loss or information entropy.

1:30:22
And this phenomenon basically states that the information

1:30:26
content of your data set will naturally decrease over time.

1:30:31
And this natural decrease is due to the human memory.

1:30:36
So when you first start out

1:30:37
with your sampling program, going out to sample your data,

1:30:41
you know

1:30:43
very well the specific and general details of your sampling protocols

1:30:48
that leads to the data that you have collected.

1:30:53
As time goes on, though, and logistics

1:30:55
change your your your environment changes,

1:30:59
your sampling protocol probably will also change.

1:31:02
And as your sampling protocol changes, you will potentially forget

1:31:07
about the specific and general details of your older

1:31:10
sampling protocols.

1:31:11
Because you're not,

1:31:12
Because you don't really need to remember that to execute the current

1:31:17
sampling protocol.

1:31:18
But that doesn't change the

1:31:19
fact that that older sampling protocol

1:31:23
was what created the older data set in the first place.

1:31:27
And so if you want to still have access

1:31:30
to the data that was collected before,

1:31:34
you still need to remember the specific and general details associated

1:31:38
with how you collected that old data in the first place.

1:31:42
The good news is to be able

1:31:45
to deal with information loss is fairly straightforward.

1:31:49
All you have to do is identify the specific and general details that are

1:31:54
required to unlock that information, and then just record it down.

1:31:58
And that process essentially makes up

1:32:01
the concept of metadata.

1:32:05
And so before we jump into metadata, I'd like to just get a sense of, where

1:32:09
everyone's at in terms of how comfortable they are with the concept of metadata.

1:32:15
So, you can follow the link that Rosie put in the chat.

1:32:20
It goes to a menti meter.

1:32:22
And if you can take a minute or two to answer,

1:32:25
these questions, then we can we can go from there.

1:32:36
Okay.

1:32:36
I don't know how you did it, Cat, but I don't.

1:32:39
I'm not getting the the live version of this.

1:32:43
Thank you. So click present.

1:32:46
I need to click present.

1:32:51
Okay. So.

1:32:52
Yeah.

1:32:54
Okay. Oh.

1:33:00
Okay,

1:33:02
So I think this is what I was somewhat expecting, that most people have

1:33:06
at least heard of the term metadata

1:33:09
and have a basic understanding of what it is.

1:33:13
Uh so hopefully I'll be able

1:33:16
to provide you all with at least a little bit more information on this

1:33:19
so that it makes you a little bit more comfortable with the concept.

1:33:25
Okay.

1:33:27
Thank you.

1:33:29
So metadata

1:33:32
is defined fairly straightforward as data about your data.

1:33:35
And there are many different types of metadata out there.

1:33:40
But I just want to highlight that what type is important is based

1:33:45
entirely on the field that you're in and the system that you're in.

1:33:50
And so for example here at IEP we collect a lot of ecological data.

1:33:55
And a lot of our data sets are long term data sets.

1:33:58
And so historical metadata is one type of metadata

1:34:02
that's extremely important for our system.

1:34:06
But overall metadata

1:34:09
will be able to provide us with critical content text

1:34:13
to help us understand, manage, find access and use the raw data that we have.

1:34:21
And so to give you all

1:34:23
a more concrete understanding of how useful metadata can actually be,

1:34:28
here's an example of a five row

1:34:32
many columns of raw data.

1:34:35
And I hope the first thing that jumps into your mind is just

1:34:38
how unusable this raw data is in its current form.

1:34:43
If I give you this data table in its current form,

1:34:47
you probably don't know what these columns are about.

1:34:49
You don't know what this data set is about, and therefore you don't really know

1:34:53
how you can even use this data set to begin with.

1:34:57
And so now let's make it a little bit more useful by introducing just one

1:35:01
layer of metadata that a lot of people don't realize is metadata.

1:35:05
And so that's the column names.

1:35:08
And so

1:35:08
here, with just this one layer of column names,

1:35:11
this data set of raw data becomes a lot more useful.

1:35:15
We realize now that there's name data here sex data, age data.

1:35:21
And when you pair that with the ticket

1:35:23
fare, cabin, embarked column,

1:35:27
then you might be able to to come to the conclusion

1:35:31
that this might be a passenger information data set of some sort of vessel.

1:35:36
Right.

1:35:37
But that can be gleaned from just one

1:35:40
layer of metadata.

1:35:43
We can add another layer of metadata.

1:35:45
And this is a variable key.

1:35:47
And these keys are really important, especially if your data set

1:35:51
uses abbreviations for your column names.

1:35:55
Like a lot of IEP surveys do.

1:35:58
And so for example the column embarked here

1:36:01
is actually the port of embarkation.

1:36:05
And then the levels within this column

1:36:08
all represents European towns.

1:36:11
And so with this next layer of metadata we can now update our mental model

1:36:15
of this data set up being passenger information data set of

1:36:20
perhaps a ship that travels between European countries.

1:36:26
Now we can add another layer of metadata.

1:36:29
And this is the variable notes

1:36:30
which gives you more specific information about your data columns.

1:36:35
Take for example the age column here it says that age is fractional

1:36:40
if it's less than one, if age is estimated,

1:36:43
it's in the form of 0.5.

1:36:46
So if I was a researcher and I was interested

1:36:49
in the age column, I now have justification

1:36:55
to remove, any data that is in the form of 0.5.

1:37:00
If I didn't want to work with estimated data.

1:37:04
Another variable of interest here is p class.

1:37:09
P class stands for the proxy of socio economic status first,

1:37:12
second or third class.

1:37:15
And so this is a fairly archaic concept.

1:37:19
So now we kind of have a time period,

1:37:22
information to stick into our mental model of what this data set actually is.

1:37:26
And so hopefully at this point across

1:37:30
these three layers of metadata,

1:37:33
you can perhaps guess what this data set is about

1:37:38
and what this data set is of.

1:37:41
Here's the last layer of metadata.

1:37:44
And it's a bulk deck diagram of where the cabins are located

1:37:49
for the first, second or third class passengers.

1:37:53
And so hopefully at this point, you were able to arrive at the conclusion

1:37:57
that this is the passenger data set of the Titanic.

1:38:03
And going through that

1:38:05
exercise, just want to highlight just how powerful these metadata layers,

1:38:11
are in making the raw data set usable.

1:38:17
And obviously the middle layer,

1:38:20
the metadata layer themselves can can tell you exactly

1:38:24
what this data set is, tell you exactly what these, the data columns are, right.

1:38:30
And like the abstract and the title of the data

1:38:33
set are metadata layers upon themselves.

1:38:36
And so I just wanted to highlight here metadata can be extremely useful.

1:38:43
And that's

1:38:43
why metadata matters so much because it allows you

1:38:47
to build a foundation that then the legacy of your data set can rest upon.

1:38:52
Right.

1:38:53
the Titanic data

1:38:56
in the Titanic sank more than a hundred years ago,

1:38:59
but because the metadata was recorded and up kept all this time,

1:39:04
we're still able to utilize that data set today.

1:39:10
Okay,

1:39:11
moving on to the second polling question.

1:39:14
Before we get into the IEP data sets, how important

1:39:18
do you think comprehensive metadata is for IEP data set?

1:39:22
So it's the same link.

1:39:24
I think you just go to the second, question.

1:39:27
If you can take a moment to answer that, please.

1:39:54
Okay,

1:39:56
so it seems like I need to find that one person.

1:39:58
That said, it's only important. I'm not extremely important,

1:40:03
but, Yeah, so

1:40:04
I agree, it's, should be extremely important

1:40:08
because it allows us directly to utilize our raw data to its fullest potential.

1:40:13
If we want to learn anything about our system.

1:40:18
Okay,

1:40:20
so moving on to talking about our IEP data sets.

1:40:24
Our data sets are actually fairly unique.

1:40:27
And, there are all long-term data sets, with the oldest one being the Summer tow net

1:40:33
beginning in 1959.

1:40:37
They're also unique in that

1:40:39
they're all ecological data sets, but they target different things.

1:40:44
So some look at phytoplankton, zooplankton, and some look at,

1:40:48
abiotic water quality and nutrient data.

1:40:52
Some look at shrimp, fish, larval fish, adult fish.

1:40:55
And so there's a lot of moving parts

1:40:58
and complexities to our IEP data set.

1:41:02
And because of that I just wanted to highlight

1:41:06
the importance of one type of metadata that is specific to our IEP dataset.

1:41:11
And that's historical metadata.

1:41:14
So historical metadata refers to the documentation

1:41:17
of the origin, modifications and movements of the data over time.

1:41:22
And again extremely important for us because of our long term nature,

1:41:27
the long term nature of our surveys, but is also really important

1:41:32
because we have a tendency to focus only

1:41:36
on documenting our current metadata, our current sampling protocols.

1:41:41
And again, that doesn't change the fact that our older protocols were

1:41:46
what created the older data in the first place.

1:41:49
And so if we wanted to understand

1:41:50
our system through time, all the data that we've collected through time,

1:41:54
we have to know that metadata for that entire time period.

1:42:00
So here is an example

1:42:01
of the yearly changes or some of the yearly changes of the SLS,

1:42:06
which is the smelt larval survey ran by the CDFW.

1:42:11
We can see here that the project

1:42:13
started in 2009, and over time,

1:42:17
additional sampling months have been added before

1:42:21
and after the original time span sampling season.

1:42:26
And then we also see that additional stations have been added

1:42:30
or removed from the Napa River and the San Pablo Bay

1:42:34
over time.

1:42:36
So let's say

1:42:38
I was a researcher interested

1:42:40
in studying the Napa River and larval fish and the larval fishes

1:42:44
in the Napa River, and I wanted to utilize the smelt larval survey.

1:42:48
It would be really critical for me to know

1:42:51
this historical metadata.

1:42:54
I might be wondering why from 2009 to 2014,

1:42:59
there isn't any catch data,

1:43:02
within the Napa River stations.

1:43:04
Same thing from 2019 to 2022.

1:43:08
Even though there's data in between.

1:43:10
Was it because sampling wasn't done or was

1:43:12
it because there were just no catches and that those stations.

1:43:18
And so that's the type of thing

1:43:20
that makes historical metadata really useful.

1:43:26
And so hopefully

1:43:27
at this point, I've at least made the case that metadata

1:43:32
is extremely important, and extremely important to our IEP surveys.

1:43:36
But where do we get started in documenting

1:43:40
rich metadata for our raw data?

1:43:44
And the key thing to know, to understand, to get started with

1:43:48
this is to recognize that

1:43:52
data sets regardless of their specific content,

1:43:56
all share common elements of metadata.

1:43:59
And so if you can theoretically create

1:44:02
a list of these standardized elements across all of our data sets, it

1:44:07
then allows you to facilitate consistency in how

1:44:11
we document our metadata.

1:44:14
And so here's an example of, a list of standardized elements.

1:44:20
You can see, it's title, PI, publication date, abstract.

1:44:25
All of these elements or theoretically

1:44:28
can be shared among all of our IEP surveys.

1:44:32
And so when you have a list

1:44:33
of these standardized elements, that's called a metadata standard.

1:44:39
And so

1:44:40
if we are able to adopt a metadata

1:44:43
standard it basically enables that, it basically,

1:44:48
increases the ease of use, for all parties,

1:44:52
for the data steward, for the data users, for the community as a whole.

1:44:58
And there are many different metadata standards out there,

1:45:02
but one that I would like to highlight is called EML

1:45:06
or Ecological Metadata language.

1:45:09
EML was first created in 1997,

1:45:13
and over the decades have had a lot of feedback

1:45:17
from the ecological, the ecology community,

1:45:22
that has essentially made it so that the language

1:45:26
is specifically curated for ecological data.

1:45:30
And so that's one really, nice characteristic

1:45:34
of the language that is applicable to our data.

1:45:39
The other characteristic is just, format.

1:45:42
And the authors of EML have opted to,

1:45:45
to use technically as XML or in the language of XML,

1:45:51
but all that is, is just a nested list

1:45:55
of your, elements and attributes.

1:45:59
And so here's an example and here's what EML actually looks like.

1:46:03
Tey're just nested list of information.

1:46:06
But because of these two characteristics, one being curated specifically

1:46:10
for ecological data and the next being the structure that it has,

1:46:15
it's actually fairly ideal

1:46:17
for describing complicated ecological data.

1:46:23
So to

1:46:23
give you a more concrete example of EML and how applicable it is

1:46:26
and how easy it is to use,

1:46:30
here is EML applied to the SLS.

1:46:33
We see here there are three layers of metadata.

1:46:37
Three levels of metadata,

1:46:39
with the first level being the name of the survey SLS.

1:46:43
The second layer is the creator.

1:46:45
The third is the organization name.

1:46:47
And so if you read this backwards,

1:46:49
you essentially get the information that you would be interested in.

1:46:52
So IEP is the organization that created the SLS.

1:46:57
And so you can then go on to add additional

1:47:01
layers of metadata.

1:47:04
Here's a more complicated layer which is the data entity layer.

1:47:09
Within this we see that this data set has a data

1:47:12
table name catch CSV, which is the

1:47:17
which is described as the fish catch data of the smelt larval survey.

1:47:21
It has a specific size on your hard drive,

1:47:25
and there are several columns within the CSV, one being the date

1:47:30
when sampling occurred and another being the station,

1:47:33
in which sampling at which sampling occurred.

1:47:37
And so I hope that you can see that because of this structure,

1:47:41
even though it sounds complicated, it's actually fairly readable.

1:47:46
And it's flexible enough to describe,

1:47:50
a whole lot of different information, using the same file structure.

1:47:56
And so

1:47:57
that's essentially the practical guidance that I have for

1:48:01
creating rich metadata within our system.

1:48:06
And that is to simply adopt EML.

1:48:09
Have this as a starting point.

1:48:11
EML is again made with ecological data in mind.

1:48:15
It's very comprehensive.

1:48:17
And and most importantly, it's

1:48:19
extendable because of the of the structure that it has.

1:48:22
And this is really important for us at IEP

1:48:25
because of the historical metadata component of our data set,

1:48:30
the long term nature of our data set that we have to account for.

1:48:33
Historical metadata isn't really as pronounced in the EML as we would like,

1:48:39
but because of the extendable nature of the, language,

1:48:44
we can simply just add this as one additional type,

1:48:49
to also document alongside everything else that EML identifies.

1:48:55
Another reason why I,

1:48:58
am picking EML or recommending EML

1:49:01
is that is the officially supported standard for EDI, which you've all heard

1:49:06
about, earlier in, in the earlier talks

1:49:09
as, the recommended data repository

1:49:13
that IEP for IEP surveys

1:49:17
to publish into and EDI is

1:49:21
a really nice in that they provide infrastructure.

1:49:25
They have infrastructure in place

1:49:27
to help guide users through this documentation process.

1:49:32
And two tools that I would like to just highlight very quickly

1:49:36
is easy EML and EML assembly line or EAL.

1:49:44
And so when you make an account

1:49:46
with EDI, you can then access easy EML.

1:49:50
Easy EML will always give you this welcome prompt which describes

1:49:55
in very in a lot of details how you can use this wizard, this tool.

1:50:01
There's also a short demo video on YouTube that will guide you through it as well,

1:50:06
if that is easier for you to use.

1:50:09
But essentially all it is, is on the left here

1:50:13
you have a list of all of the metadata

1:50:17
elements that you are expected to fill out

1:50:21
and within each of these

1:50:24
elements, if for some reason you are confused

1:50:28
about what you need to, record within each of these fields,

1:50:33
they have these question marks all over the place, and you can click

1:50:37
on those question marks to get additional details and additional help

1:50:40
as to what they are looking for, for you

1:50:43
to fill out for this element.

1:50:47
Another useful feature, are these checks

1:50:52
that occurs in the background as you are filling out your easy EML

1:50:57
metadata documentation.

1:51:00
So essentially, as you're writing through and filling out all this information,

1:51:05
if at the end of this for some reason, one of these checks are still read,

1:51:11
you can simply click on them and it'll tell you

1:51:14
exactly what you're missing or what you still need to fill in.

1:51:17
And so I find that really helpful as basically

1:51:21
a guide to make sure that you're on track.

1:51:24
And then the last feature that's really nice

1:51:27
is if you then decide to publish to EDI,

1:51:31
you can submit the package to the data curation team from EDI.

1:51:36
And that essentially allows you to tap into

1:51:41
a team that have seen thousands of different data publications.

1:51:46
And they can serve as a second pair of very experienced eyes on your data set

1:51:52
before it gets published to the public.

1:51:55
And so there's a lot of nice things going for

1:51:59
easy EML.

1:52:03
EML assembly line, or EAL,

1:52:05
does basically the same thing as easy EML and creating that rich metadata,

1:52:10
but it's geared more towards users that know how to use R

1:52:16
because everything is done programmatically through R.

1:52:19
And so if you have experience working in R,

1:52:21
and it's really important for you to automate your data publication process

1:52:28
then I would

1:52:29
recommend checking out easy or EAL here.

1:52:33
and it's available on the GitHub page.

1:52:36
And you if you, need

1:52:41
the DUWG has also worked with Colin Smith, who is one of the authors at

1:52:45
EDI of this package, to create a guide for you to, to work through.

1:52:52
And this is on the IEP website under the DUWG tab.

1:52:58
And then for some reason, easy

1:53:00
EML or EAL don't work for you

1:53:03
then,

1:53:05
we have also created these, word documents

1:53:08
that are essentially embedded in your templates that identify

1:53:12
your metadata elements and define them.

1:53:15
And so you can just go through and follow this template and also

1:53:20
fill this out to, record

1:53:22
rich metadata for your raw data.

1:53:28
Okay.

1:53:30
Here is the next question

1:53:33
before we jump into data publication,

1:53:37
which tool can you see yourself using to document rich metadata?

1:53:41
And so I'm curious to see.

1:53:45
What the answers are here.

1:53:56
And this doesn't have to be what you're currently using.

1:53:58
Now this can just be,

1:54:01
if you're open to checking it out.

1:54:35
Okay.

1:54:36
Seems pretty split, actually.

1:54:39
And I'm actually surprised there are so many people using EAL.

1:54:43
It might just be the audience that decided to come to the,

1:54:47
to this workshop.

1:54:52
Okay,

1:54:54
cool.

1:54:54
yeah.

1:54:55
If you have a chance, I would.

1:54:57
I would say ez

1:54:58
EML is a pretty nice tool, even if you are fairly advanced.

1:55:02
And then you can take that knowledge and apply

1:55:05
it to EAL, fairly easily.

1:55:17
Okay. So

1:55:19
moving on to data publication.

1:55:22
So data publication is the process of making our research data publicly

1:55:27
available to facilitate the discovery,

1:55:30
access, use and reuse of our data.

1:55:34
I have here a diagram of the,

1:55:37
data lifecycle created by the IEP,

1:55:41
and I just wanted to highlight here

1:55:44
the data publication is the last step within this process.

1:55:48
the share, the shared point here.

1:55:53
But that there's

1:55:56
a lot of work that is involved.

1:56:00
Before you get to this last share step, there's work that has been done

1:56:04
in design, the collection, the processing of your data.

1:56:09
And so essentially, you don't want to miss this step.

1:56:14
And you don't you I would argue that you want to pay

1:56:17
attention most to this step,

1:56:21
because is

1:56:22
the culmination of all the work that you put in so far.

1:56:26
And so if you are able to publish correctly, follow

1:56:29
best practice for best practices for data publication,

1:56:33
there are benefits to everyone.

1:56:34
There's benefits to you as the data steward,

1:56:37
to your data users, to the community that you're in.

1:56:41
but I just like to highlight some benefits to the data stewards.

1:56:45
It enhances the transparency of your process,

1:56:49
if your users know exactly what you're doing to make the raw data

1:56:53
available to them, they can interpret that data more correctly.

1:56:58
It increases the visibility, recognition, and impact

1:57:03
of all the work you do to collect and work up the data.

1:57:08
It improves the data management system that you have.

1:57:10
If everything is laid out, then you can simply just rely on that

1:57:15
built infrastructure from year to year to consistently publish your data set.

1:57:22
And then

1:57:22
ultimately it promotes data preservation.

1:57:25
Even though it might seem difficult to

1:57:29
to delete a file to delete a database, people don't realize

1:57:34
just how fragile, digital data is.

1:57:39
Your server can just go down, the file can be mislocated.

1:57:43
There can be changes made from a user.

1:57:46
and so digital file is actually very fragile.

1:57:50
And so having it preserved in multiple different places

1:57:53
is always a good thing.

1:57:56
But how do we get started on, digital publication.

1:58:01
How do we do this correctly?

1:58:03
And one way is to follow the FAIR principles.

1:58:06
And Cat touched on this earlier.

1:58:09
But the FAIR principles are guidelines

1:58:12
that are meant to enhance the findability accessibility,

1:58:16
interoperability and reusability of your data set.

1:58:20
These guidelines were developed by an international team

1:58:24
of researchers in 2016, and it has essentially

1:58:28
became become the gold standard for the data publication industry.

1:58:34
And I'm not going to go over the principles

1:58:36
here today, but you can go to their website to find out more.

1:58:40
But essentially, the guidelines all revolve around

1:58:44
the fact that you should document your metadata thoroughly,

1:58:48
to make it as usable as possible. You should adopt a metadata standard

1:58:52
to help you do step 1.1 here,

1:58:57
but also increase the consistency of your documentation process.

1:59:03
And then the third is to also lean on accepted infrastructure.

1:59:07
This isn't a new problem.

1:59:10
This has been a problem that people have been trying to solve for a very long time.

1:59:15
And you can essentially just lean on the work

1:59:19
and the progress that they have made so far.

1:59:23
And so, all that points to back to EDI.

1:59:27
Again, EDI is a data repository that IEP recommends for their surveys

1:59:33
to publish into, but they have a lot of existing

1:59:37
infrastructure in place to support their data.

1:59:42
They have things like a unique

1:59:44
and persistent DOI for each of your data package that you publish to.

1:59:48
This helps your users find your data set using that DOI site.

1:59:52
That data set. It has, it uses EML,

1:59:56
EML as its standard to produce very rich metadata.

2:00:01
It's part of a data, a data federation,

2:00:05
which means that if your data users don't use EDI,

2:00:09
they can still find your data through this federation.

2:00:14
You can access your data,

2:00:16
and metadata through any browser, through any

2:00:20
and through and through the API, which means that you don't need

2:00:24
specialized software in order to get to the data.

2:00:27
In order to access the data.

2:00:30
Obviously they provide guidance and support that I think

2:00:33
is really critical to this process through easy EML, EAL,

2:00:38
assessment reports, dedicated curation team.

2:00:41
And then finally, they also support version control of your data set each time

2:00:46
you publish, a version of your data set that is logged on their database.

2:00:51
And you can theoretically go back and access that version if you need to.

2:00:56
And then most importantly, data preservation

2:00:59
is a mission statement of the repository.

2:01:05
And this is something that people

2:01:08
might underestimate how important it is.

2:01:11
If you for some reason publish to let's say GitHub,

2:01:16
that is a third party entity.

2:01:19
It's a private company owned by Microsoft.

2:01:21
And even though it's very unlikely, Microsoft could theoretically

2:01:25
just pull the plug on that website on that service overnight and therefore,

2:01:30
you would lose your data publication and the history of your data publication.

2:01:34
and like I said, even though that seems very unlikely,

2:01:38
this has happened in the past to things like Google Code.

2:01:42
And so an entity as large as Google and have in the past

2:01:46
just pull the plug, and the users would then have to migrate

2:01:49
and scramble in order to preserve their data,

2:01:53
at that point.

2:01:54
And so data preservation is actually a really important thing

2:01:57
to consider when choosing where you want to publish your data to.

2:02:04
But all in all,

2:02:06
FAIR and EML

2:02:09
and EDI, these are all great initial steps,

2:02:13
I would say in terms of, documenting

2:02:17
your metadata, making your raw data

2:02:20
as usable as possible, publishing your, your, your data.

2:02:24
But there are things that we need to consider that are system specific,

2:02:29
to IEP, to our data set,

2:02:32
things like the long term nature of our data set, things like require

2:02:36
data portals that we have to publish to them by our funding agencies.

2:02:41
Reporting frequencies that we have to abide by.

2:02:45
And most importantly, the existing infrastructures that we have in place.

2:02:51
A lot of us have our data in Excel or an access database.

2:02:56
And so if we wanted to update that workflow to incorporate,

2:03:01
let's say, R into it, we have to account for this,

2:03:06
this change, within our

2:03:08
workflow, we can't just move everything to a different type of database.

2:03:14
We also have scripted QA/QC queries

2:03:16
that were developed a long time ago.

2:03:20
and so if we start changing things like

2:03:23
column names, in order to conform with,

2:03:26
these new standards, it might break

2:03:30
various queries that are fairly interlinked

2:03:34
with one another, within our current grammar structure.

2:03:37
And so that's also something we need to consider.

2:03:40
And so these principles FAIR, EML

2:03:43
and EDI, there are just great initial steps,

2:03:47
but we do need to pay attention to needs that are specific to our own system.

2:03:54
And that's where the data publication working group steps in.

2:03:58
And so this is a subgroup of the DUWG

2:04:02
that focuses on providing guidance that is more specific to our system.

2:04:08
And, I'm the chair of this group.

2:04:10
And we have two main functions.

2:04:13
The first is to develop guidance for the community

2:04:17
in terms of data publication, metadata, documentation, things like that.

2:04:21
And we do this by creating things like guidelines, templates,

2:04:25
guides on how to do all this stuff.

2:04:28
And we've already put in a lot of work in regards to this.

2:04:33
All of our products can be found on

2:04:37
the IEP website under the DUWG tab here.

2:04:42
So that's the first portion

2:04:45
of our main function.

2:04:48
The second, main function that we have is to actually serve as an open forum

2:04:53
to help IEP surveys through this data publication process.

2:04:58
And so essentially,

2:05:00
we hold meetings every month to work on, the first function.

2:05:05
But if any IEP surveys would like help

2:05:09
with, their publication process, then they can simply reach out to us

2:05:15
and we can dedicate one or more of those meetings to

2:05:19
provide you with tailored guidance on what you need to do

2:05:24
to improve your data publication process, things for you to consider.

2:05:28
And we'll even review a draft of your, data publication package itself

2:05:33
if we do it to EDI and see if you have everything in order.

2:05:38
And so hopefully this in addition to EML, FAIR principles,

2:05:44
EDI will round out this entire process and help everyone

2:05:49
through this entire process, through the state, the publication process.

2:05:55
Okay, so last polling question here.

2:05:58
Before I go into the last topic of culture,

2:06:02
which data repositories

2:06:04
do you prefer using for your data publication?

2:06:09
And this is more of a curiosity than anything about what

2:06:14
we are currently doing within our system.

2:07:05
It's kind of exciting to see which one would win.

2:07:08
yeah.

2:07:10
This was what I was kind of expecting.

2:07:13
That we actually do have a lot of in-house networks that we

2:07:17
publish to, and

2:07:22
that's usually pretty easy

2:07:24
to find if you're using, like, an FTP website and things like that.

2:07:27
And but at the same time, when we're all using different networks,

2:07:33
it does make it a little bit difficult if we want to

2:07:36
look across the entire system.

2:07:39
Okay, cool.

2:07:42
I'll thank

2:07:42
you for entertaining my my Menti meter there.

2:07:46
Those questions.

2:07:48
Okay, so moving on to the last topic here today.

2:07:53
I just wanted to talk about the culture

2:07:56
that we, we have and the culture that I think we need

2:07:59
to develop and embrace if we want to be successful here.

2:08:04
I think the most important thing that we need to do is to embrace

2:08:08
this transition from being data producers to data stewards.

2:08:13
I think we need to value the fact that,

2:08:17
the extra work that needs to be put in

2:08:20
to make our raw data as usable as possible is worth it.

2:08:25
And so there are two main parts to data stewardship in my mind.

2:08:28
The first is creating the technical guidelines

2:08:31
and the technical documentations

2:08:32
to guide our users on how to use our data correctly.

2:08:36
But the second is the want, the attitude,

2:08:40
the behavior that the want to put in this extra work.

2:08:45
And hopefully

2:08:47
I convince you that it's worth doing,

2:08:52
that we should really do care

2:08:54
a lot about making our data the best available data

2:08:59
that's out there, because best available data will inform best available science,

2:09:04
which ultimately informs how we manage our system.

2:09:08
And so if you're like me, who have gone into this field

2:09:13
wanting to make a difference in our natural resources,

2:09:16
I hope that you realize the data that we collect actually

2:09:21
is this the starting point

2:09:24
which we can accomplish that, making

2:09:26
meaningful changes to our system and to our field.

2:09:31
Another point that I want to make is that the future is essentially infinite.

2:09:36
And so even if our data is not currently used now,

2:09:40
there are the potential for use of our data is also infinite.

2:09:46
There can be someone later down the line that comes into our system

2:09:49
that figures out how they can use our data in a different way,

2:09:52
beyond the original scope of what we want to use the data for.

2:09:58
there can be new

2:09:58
technologies that gets, that are founded,

2:10:03
that can then unlock the use of our data.

2:10:07
And so this, this potential for use is, I hope,

2:10:12
an exciting thing for you to think about, when it comes

2:10:16
to making our data as usable as possible for as long as possible.

2:10:20
And so ultimately, this leads to the fact that each line of useful data

2:10:25
that we create, is essentially a part of a legacy that we can bill

2:10:31
to leave behind for when we decide to move on

2:10:35
from our position, our roles.

2:10:37
If we do a really good job at documenting our metadata

2:10:42
and making our role data as usable as possible, then the information content,

2:10:46
the usefulness of that data can persist for as long as possible through time

2:10:52
to reach as many people as possible, through time to be.

2:10:56
And so

2:10:57
I think it's

2:10:58
really worth it to to put in this extra effort.

2:11:02
And so how can we start then?

2:11:05
Well, one, we can obviously prioritize documenting

2:11:08
comprehensive metadata for all of our IEP data set.

2:11:12
And we can do this by adopting

2:11:15
EML or a, an equivalent standard.

2:11:19
But I recommend EML.

2:11:21
We can explore publication through the EDI repository.

2:11:26
and for those of you that are required to publish

2:11:30
to a specific repository, I would say

2:11:34
also try publishing a as well through EDI,

2:11:38
because of all the great infrastructure that it has,

2:11:43
for you to lean on that infrastructure, I think,

2:11:46
it makes the data publication process a lot smoother and a lot more robust.

2:11:51
And, I believe ez

2:11:54
EML is actually fairly easy to use.

2:11:58
Another thing you can do is to engage with the data publication working group.

2:12:03
To both follow and develop the best practices

2:12:07
that we have for working in our system.

2:12:11
we have a lot of members there that have published

2:12:14
a lot in the past, but that doesn't mean that we know everything.

2:12:18
and so if you want to contribute, I would welcome you to join the group

2:12:22
and help develop these best practices,

2:12:26
for community.

2:12:28
And then lastly,

2:12:30
to develop and embrace a culture of data stewardship,

2:12:35
to understand that it is through the data that we collect, the work that we do,

2:12:40
that a lasting legacy can be left behind

2:12:44
using the data that we collect, our research data.

2:12:47
And hopefully, I've made a good case for that.

2:12:52
So that's it.

2:12:53
thank you all for your time.

2:12:56
I think we're pretty early, but

2:12:59
thank you.

2:13:05
Yeah.

2:13:05
Thank you so much, Trinh.

2:13:08
I think you're getting a lot of applause.

2:13:10
We are early.

2:13:13
and so I think we will break early, but I did think

2:13:20
since we have a little extra time, we could take some questions

2:13:25
for the first portion of the,

2:13:29
today's workshop, if there are any.

2:13:32
I also wanted to.

2:13:33
I just made an extra menti slide.

2:13:35
since those are fun for me, and I'll stick that in as well.

2:13:40
Just spend five minutes,

2:13:43
answering

2:13:43
that and seeing if there are any questions.

2:13:47
I haven't seen any in the chat that haven't been answered yet,

2:13:53
but let me put the.

2:13:56
Menti back in.

2:14:00
So there's anything that kind of, came up.

2:14:03
You might not know all of the existing guidance, templates, code,

2:14:09
but if there's anything that you have wanted,

2:14:12
additional guidance on or, various data sets that

2:14:16
haven't been published, just feel free to stick those in.

2:14:19
and then if there are any questions, feel free to raise your hand or stick

2:14:22
it in the chat.

2:14:27
I also want to mention that we will

2:14:31
be sending out, a survey

2:14:34
at the end of the training today, so I'll probably ask this question again.

2:14:40
Give you time to think about it,

2:14:43
but hopefully I'll have time to give some more feedback

2:14:46
for future workshops.

2:15:15
And if anyone in the audience knows of resources

2:15:18
for the things that are popping up on the screen, feel free to

2:15:23
send that to me so I can add it or,

2:15:25
put it in the chat.

2:15:31
I know that there was,

2:15:34
work looking into digital data sheet,

2:15:41
software and

2:15:49
we have I think we have that in our resources.

2:15:52
I'm not sure if it provided any information

2:15:55
for integrating with access.

2:15:59
And then

2:16:00
I think there is guidance for linking EDI

2:16:03
to see CNRA.

2:16:07
Rosie, do you know where that lives?

2:16:11
There is guidance.

2:16:13
Oh, sorry.

2:16:14
There is guidance, up on the DUWG site right now that Dave and I are like

2:16:20
prepping a updated version

2:16:22
of it to, you know, be done next week or so.

2:16:26
because the old one was a little bit old.

2:16:28
and if anyone wants to test out the new version

2:16:32
and see how it goes, we'd love volunteers.

2:16:37
Great.

2:16:42
I just put the link to that in the chat.

2:16:46
Thanks.

2:16:48
And then I think for the one about

2:16:52
data flags outliers in QC.

2:16:58
Rachel might have mentioned

2:17:00
a package that deals at least with,

2:17:04
flagging continuous water quality data,

2:17:09
but Dave and some others have developed so

2:17:15
maybe Dave can put the link for that in.

2:17:17
But,

2:17:20
it doesn't necessarily handle

2:17:21
all data types, but does have some,

2:17:24
some types of tests that you can consider for other data types.

2:17:28
Yeah,

2:17:29
it's for both continuous and discrete collected data.

2:17:33
But yeah I'll put it in the chat. Okay.

2:17:35
Thanks. Sure.

2:17:39
And I think is there a group that you are

2:17:41
that's coming up with other kinds of guidance around this?

2:17:46
Yeah.

2:17:46
So I mean, that's kind of that R package is kind of being developed

2:17:50
by that group.

2:17:51
and it's yeah, it's a kind of an internal DWR,

2:17:55
group called the Outlier Detection Working Group.

2:17:58
And so, Rachel already kind of shared, a few things from that group,

2:18:03
including the, guidance document that we created for both discrete

2:18:07
and continuous, outlier detection methods.

2:18:11
So you can take a look at that to.

2:18:14
Okay. Thanks.

2:18:16
Sure.

2:18:20
For Tableau I know that USGS is exploring

2:18:24
Tableau as a platform for visualizing some of their data.

2:18:28
So, maybe as that gets developed

2:18:31
we can explore that a little bit more.

2:18:34
I don't know too much about power BI.

2:18:39
but yeah, I think that would be good

2:18:40
to look into what else is out there and what can

2:18:44
visualize data real quickly.

2:18:48
You will hear about SacPASS and,

2:18:51
Bay-Delta live this afternoon.

2:18:53
So those are some

2:18:55
portals that

2:18:58
pull in

2:19:00
data that get updated, like Denise mentioned

2:19:03
in real time or, in other frequencies, weekly, monthly, things like that.

2:19:08
so they kind of pull up data as it comes out

2:19:12
and have preset visualizations.

2:19:17
So those are some options for people

2:19:20
wanting to have set visualizations

2:19:23
for their data.

2:19:33
But looks like we're slowing down on the suggestions.

2:19:37
Feel free to keep thinking and sticking answers in.

2:19:40
I can I can look at it later.

2:19:41
I don't have any other Menti meter

2:19:47
polls coming up, so we'll keep this open.

2:19:50
Maybe we can revisit later.

2:19:53
And like I said, we'll send out more questions later too.

2:19:56
But I want to let

2:19:59
So this is to bring in some of those folks that might not be familiar,

2:20:03
provide a little background, on the Yolo Bypass in general.

2:20:06
but of course, with these tabs, you could skip past that

2:20:08
if it's not of interest to the reader,

2:20:12
and then dive into the program itself.

2:20:15
And I'll talk about this more in a little bit.

2:20:16
But these are some of the really cool interactive maps

2:20:18
that you can add in with ArcGIS story map.

2:20:20
So it has this pop up legend.

2:20:23
and then you can scroll in and scroll out and use your mouse to navigate around.

2:20:29
so I really liked this feature, and I'm not super map savvy,

2:20:33
and it was fairly easy to put this together.

2:20:36
I'll dive into that

2:20:37
a little more with an example of what starting one of these would look like.

2:20:41
then highlighted all of our different methods with photos

2:20:45
and put some different links to some of our outreach.

2:20:49
initiatives and then references at the bottom.

2:20:52
So pretty straightforward and simple

2:20:54
way to try and capture

2:20:56
the main points of our program that I wanted to share with folks.

2:21:04
I want to provide

2:21:04
one other example, which is of a River reborn, which is about,

2:21:10
the Elwha River restoration that happened in the state of Washington.

2:21:13
And you'll notice this was created by ESRI Story maps teams.

2:21:17
So this is an example of how fancy you can get with it.

2:21:21
and their goal was to document the return of salmon

2:21:24
after a major dam removal.

2:21:27
So I'm going to flip through this fairly quickly.

2:21:29
But just to give you a flavor of the more advanced capabilities

2:21:33
that you could explore with Arc Story maps.

2:21:37
So here you've got these scrolling bars, you've got a map that comes in,

2:21:41
and then you zoom into it

2:21:42
and it takes you through different points in the river system.

2:21:46
So depending on what you're trying to get across, this

2:21:49
the map features are much more substantial than what I showed in my YBFP.

2:21:53
For example, you can see the restoration.

2:21:56
They have some really awesome visuals.

2:21:59
and I include this link at the end and I'll throw them in the chat

2:22:02
after my presentation.

2:22:04
So you can explore this more if you'd like to check it out.

2:22:07
It's a really neat story map.

2:22:12
So to

2:22:13
get started with Story Map, you'll want to outline your story goals.

2:22:16
Determine your account type.

2:22:17
So there I'm pretty sure I made the YBFMP one

2:22:20
with a free ArcGIS account, and it allows you to do a lot.

2:22:24
but there are even more features that are available if you have a creator

2:22:28
or GIS professional user type, which might be accessible through your agency.

2:22:34
and then just know that there are a ton of different online

2:22:36
resources that you can access.

2:22:38
So I'll provide links to that as well.

2:22:40
But it's really easy to find tutorials, and information to get you started.

2:22:47
But I'd like to really just hit home how easy it is to get started

2:22:51
by showing you what it would look like to start a map.

2:22:54
So this is just in my profile for ArcGIS Story maps.

2:22:58
You can see that Yolo Bypass one is here.

2:23:00
but if I go and I click New Story, I want to start from scratch.

2:23:06
It launches the story builder

2:23:09
and you can add a title,

2:23:16
And then you can start adding in,

2:23:19
content right away.

2:23:20
So there's options for text buttons.

2:23:23
That's like how I created some of those links.

2:23:25
And the Yolo Bypass one.

2:23:27
you can put in tables, images

2:23:30
and it's pretty easy how they make the images work.

2:23:33
So I'll just give a quick example. Okay.

2:23:37
So once you add an image there's all these simple ways that you can format it.

2:23:41
So you can have it on the side.

2:23:42
You can have it in the middle.

2:23:44
so they have a lot of these preset functions.

2:23:47
So you don't have to sit there and tweak like you would in say a PowerPoint

2:23:51
or something like that.

2:23:52
There's a lot of preset functions that make it really easy to throw content in.

2:23:57
You can add in videos,

2:24:00
there are some really cool immersive options.

2:24:03
So there's this sidecar feature and you can see they have docks.

2:24:07
So that would be as you're scrolling down

2:24:09
things are moving along with you floating.

2:24:12
Or you have these little floating text boxes.

2:24:15
and I've seen some really cool ones that have utilized this.

2:24:18
Hailey Mica, one of my colleagues at DWR,

2:24:21
she did a story map for the Wallace Weir Fish Rescue facility.

2:24:26
That's actually what inspired me to make my story map

2:24:28
and she had utilized some of these cool floating features,

2:24:32
which I think work really well,

2:24:33
especially if you have a smaller amount of information to put across.

2:24:37
I found it

2:24:38
didn't work as well

2:24:39
with as much information that I wanted to put in the YBFMP one,

2:24:43
but it's an option depending on what the goals of your story map are.

2:24:47
There's also these slideshow functions, so you can use those if you'd like.

2:24:52
and then they also have a map tour, which is kind of

2:24:56
like what we saw with the Elwah River example.

2:24:59
I'm not going to get into that one right

2:25:02
now, but I will show how easy it is to embed a map.

2:25:05
So if you click on map, if you have an ArcGIS online account

2:25:11
that you use for mapping, you can access your maps through here

2:25:15
by adding them in, but if you don't already

2:25:19
have a map created and say you don't, you're not super GIS savvy.

2:25:23
you can just go

2:25:24
click New Express Map and it'll take you to this really,

2:25:29
like well-executed, user interface

2:25:32
where you could say, scroll in here, you have the San Francisco Bay.

2:25:37
let's say you were working in the marsh,

2:25:40
and you just wanted to create a quick map of where your study sites were.

2:25:44
So let's say you had three

2:25:47
different sampling sites along Montezuma Slough.

2:25:52
You can plug those and say you want to show,

2:25:55
you know, the different path that you took or something like that.

2:26:00
You can add that in.

2:26:01
You don't need all the coordinates and everything

2:26:03
if you don't have those on hand.

2:26:05
Just for a quick visual, say you want to circle an area of interest.

2:26:09
you want to add a little text.

2:26:16
It's very user friendly.

2:26:17
and at that point you can just click save

2:26:20
and it plugs it right in and then you've got that easy.

2:26:24
navigable map right there in your story map.

2:26:29
So you can keep playing around with these things, add in what information you need,

2:26:32
and then you would just go, you can preview it.

2:26:34
You can publish it.

2:26:36
And it's really straightforward.

2:26:40
And with that, I have some links

2:26:43
that hopefully can help folks get started and provide some,

2:26:47
materials or inspiration through some of the other story maps that have been made.

2:26:51
I will go ahead and link those, and I hope this inspires you

2:26:55
to try and tell your project story through, ArcGIS Story map.

2:27:01
Thanks.

2:27:04
Thank you so much, Nicole.

2:27:06
That was a great demo.

2:27:08
So yeah, hopefully everyone's going to go make their own story

2:27:11
map now.

2:27:14
We have up next, Sam Bashevkin.

2:27:18
from the Water Board

2:27:21
and he's going to present on R shiny.

2:27:26
Great. Thanks, Cat.

2:27:27
And hi, everyone.

2:27:28
as Cat said, my name is Sam Bashevkin, and I'm with the State Water Board

2:27:33
and I'm going to attempt to give an introduction

2:27:37
to shiny and 15 minutes.

2:27:40
It'll just be, scratching the surface sort of talk, but,

2:27:43
hopefully you can use it as inspiration

2:27:46
to dig into a deeper.

2:27:50
So first, what is shiny?

2:27:52
It's an R package to build interactive apps with a graphical user interface.

2:27:58
You can use basically any R functionality,

2:28:00
including external packages within the apps.

2:28:04
They can be either hosted online,

2:28:06
and distributed as web pages, or used locally.

2:28:12
Why would you want to use shiny?

2:28:14
I think the main reason is to bring R functionality to non R users.

2:28:20
and you know, related to

2:28:22
that would be to communicate science.

2:28:26
But it can also be helpful,

2:28:28
you know, even if you are an Ruser, to build something that'll let you

2:28:31
quickly toggle different parameters and visualize the effects of it.

2:28:34
So it can be a neat way to just build yourself a tool

2:28:38
to visualize data.

2:28:41
But why would you not want to use shiny?

2:28:43
The main reason is if you're trying to reduce your frustration load.

2:28:46
Shiny is not, a good thing to dive into.

2:28:50
If you're trying to just show off your impressive

2:28:52
coding skills, you know, there might be better ways to do that.

2:28:57
And if you're trying to build something

2:28:58
you want to have time to maintain, Shiny is also not the best for that.

2:29:02
because you always need to be maintaining

2:29:05
something like a shiny app.

2:29:09
So I'm going to first walk through a basic

2:29:12
shiny app and what it looks like and how to start it.

2:29:15
So if you want to start a shiny app from within R studio,

2:29:20
you just click the little plus icon and then there's an option for Shiny web app.

2:29:25
And that will build you this app,

2:29:28
which is just plotting some data

2:29:30
from the Old Faithful geyser and Yellowstone National Park.

2:29:35
in a histogram.

2:29:36
And you can just choose how many bins the histogram is,

2:29:39
distributed by.

2:29:42
So this is what the code behind that app looks like.

2:29:46
And there's two main components outlined in the black rectangles.

2:29:50
The top part is called the user interface.

2:29:53
The bottom part is the server.

2:29:55
And they talk back and forth to one another.

2:29:58
So the user interface will send your user inputs.

2:30:01
So like the number of bins you select down to the server.

2:30:04
And then the server will produce a graph and send that back

2:30:07
up to the user interface so that the user can see the graph.

2:30:13
So an example of how that works.

2:30:15
On the top I have this slider input object surrounded in the black box.

2:30:20
And you can see on the right it's just a picture of what that produces.

2:30:24
and that's, you know, how you select how many bins you want.

2:30:27
And then that information gets sent over in this object called input$bins.

2:30:32
And that's how the server, receives the information

2:30:35
and uses it in the graph.

2:30:38
And then in the reverse

2:30:40
order down here in the server section

2:30:43
we have this production of this histogram, which is produced.

2:30:48
And then it's sent back up to the user interface.

2:30:51
and it produces the graph that you can see on the side.

2:30:57
So that's just a quick example of how they kind of talk back

2:30:59
and forth to one another.

2:31:03
So just diving a little bit deeper into how Shiny work.

2:31:07
So the user interface portion, which I'm showing on the screen now,

2:31:11
is analogous to a list of user viewable components that are wrapped

2:31:15
in different functions to set the layout.

2:31:20
So actually, before the user interface is even defined,

2:31:24
there's an area where you can load things that are used by the entire app.

2:31:27
So this is where you load packages, functions or anything like that.

2:31:31
and then within the user interface I've surrounded in these yellow

2:31:35
rectangles the functions that you can use to control the layout.

2:31:38
So how you know, items are distributed across the app page.

2:31:43
So in this case it uses this fluid page

2:31:46
function, to define the overall page layout.

2:31:49
There's a title panel that has the title, a sidebar layout and sidebar panel

2:31:54
that have the, the user input, and then the main panel

2:31:58
with the plot output that produces the plot.

2:32:00
So there's a bunch of these functions, and you can mix and match them

2:32:03
to, control how you want the app to look.

2:32:09
And then for the

2:32:09
server portion that's analogous to a series of functions

2:32:13
with the inputs from the user interface, as the arguments

2:32:16
of that function.

2:32:19
So then just going back again,

2:32:21
this is what the app would look like they produce.

2:32:24
So what you're seeing here is everything that the user interface tells you to see.

2:32:28
But then the graph itself is being produced by the server

2:32:31
and sent back up to the user interface.

2:32:36
So I'm going to go just a little bit deeper

2:32:38
to show some examples of more complex functionality that you could use.

2:32:43
So this is a slightly more complex app.

2:32:46
You'll notice I added one more thing.

2:32:48
You can select that.

2:32:49
You can select the column from the data set to plot.

2:32:52
And then there's a button here for run.

2:32:53
Those are the two things that I've added.

2:32:55
And I'll show you kind of why and what those do.

2:33:00
So stepping back a little bit, there's this important function

2:33:04
in shiny called reactive that can create intermediate values

2:33:07
that will be triggered by any change in their dependencies or inputs.

2:33:12
So it's a way you can basically define a function.

2:33:14
And then whenever a user selects something different,

2:33:17
it'll trigger that function to be rerun using the new input.

2:33:21
But if you don't want everything to update

2:33:22
as soon as the inputs change, then you can use event reactive.

2:33:27
And that's useful when the user has to choose multiple inputs

2:33:29
and you don't want to process the process to trigger an update until they have

2:33:33
made all of their selection.

2:33:34
So if they've to select from, you know, 50 different,

2:33:38
questions, you don't want the graph maybe to update

2:33:41
after they answer every single question. Maybe you want them

2:33:44
to answer all the questions and then they can say update graph.

2:33:47
And that's why I added that run button in the in the app.

2:33:53
So here

2:33:55
is the same app from before.

2:33:57
But surrounded in the black boxes are the additions that I've made.

2:34:01
and I'll walk through in the next

2:34:03
few slides kind of what the reasoning is for these, but you see, there's

2:34:07
the radio buttons are the column to plot the action button is that run button.

2:34:11
And then I had to alter the plot a little bit to take those inputs.

2:34:16
So in the user interface, just zooming in on this addition,

2:34:20
we have this radio button, as I said, column to plot.

2:34:24
And that produces the little buttons that popped up above.

2:34:29
And then the action button produced that run, the run button.

2:34:33
So it's, you know, it's pretty simple to add additional inputs.

2:34:36
It's really just, you know, a line of code.

2:34:41
And then in the server, this is where I'm using

2:34:43
this event reactive function that I mentioned earlier.

2:34:46
So this is saying it's telling the app to watch this run button.

2:34:51
And whenever the run button is clicked, regenerate the plot.

2:34:55
But if anything else happens you don't regenerate the plot.

2:34:57
Only when that run button is clicked.

2:35:03
So this is just,

2:35:04
you know, showing graphically what I was just saying.

2:35:07
The run button is connected to this input dollar sign run.

2:35:11
object.

2:35:12
And that's what is triggering the block to be regenerated.

2:35:18
And again, this is just what that plot looks like.

2:35:20
So now you can select the number of bins.

2:35:22
You can also select the column you want to plot.

2:35:25
And then you can control when the graph is updated.

2:35:27
So if you change some of these and you don't

2:35:29
click run, the graph is not going to change.

2:35:31
When you do click run, then the graph update.

2:35:36
So debugging shiny apps are hard to debug.

2:35:39
That's where a lot of the frustration I mentioned earlier comes from.

2:35:45
I'm just going to talk through some general high level recommendations

2:35:49
I have that have helped me get out of some tricky situations.

2:35:53
The first step

2:35:55
that helps is always to start by running the code outside of the app.

2:35:59
It's a lot easier to try to isolate the issue.

2:36:01
If you pulled the relevant code out of the app and then try running it

2:36:04
and see why it's failing, rather than trying to click through the app,

2:36:08
and see when it crashes.

2:36:13
But then many bugs that I've been unable to replicate outside

2:36:16
the app have been related to an improper sequence of events.

2:36:20
and this is an issue that at least I've encountered a lot.

2:36:25
And I'll just explain what I mean.

2:36:27
So generally with an app you want, first the input is selected,

2:36:32
then the data are processed, and then the graph is produced.

2:36:35
That's like the logical sequence of how things would occur.

2:36:39
But you know, shiny doesn't know how to do things in this order.

2:36:44
You have to tell the app to do things in the right order.

2:36:46
So if the ordering is wrong for some reason, it might

2:36:51
get the input and then try to produce the graph before the data been processed.

2:36:54
And that will, you know, create an issue because it's not going to be

2:36:58
receiving the data that it expects.

2:37:01
So that's just something I'd recommend looking out for.

2:37:04
Some ways to fix it.

2:37:05
There's a function rec, that can be used to ensure

2:37:09
that processes don't trigger until their dependencies exist.

2:37:13
So you can use it to kind of explicitly control

2:37:15
the flow of the process.

2:37:19
And you can also set default values or use if analysis

2:37:22
or similar functions to try to control how the app is functioning.

2:37:30
So there's a lot of add on packages for shiny

2:37:33
or for visualization in general that are that I found very helpful to use

2:37:37
in shiny apps.

2:37:39
First there's the ggiraph or plotly,

2:37:43
packages that allow you to make coverable interactive graphs.

2:37:48
I use ggirpah, but I know a lot of other people use plotly.

2:37:53
There's the Shiny Widgets app that has more and prettier,

2:37:56
input types that are nice to use.

2:38:00
Leaflet can be used to produce interactive maps

2:38:03
like the one on the right here and leaflet.

2:38:06
minicharts can be used to produce pie and bar charts over maps.

2:38:10
Also used in the map to the right.

2:38:13
And then you can also customize the app by inserting HTML or CSS

2:38:18
if you're familiar with those, or if you're comfortable googling

2:38:21
and pasting and seeing what happens, which is how I usually do it.

2:38:28
So then if you have an app, what do you do with that?

2:38:31
Normally you'd want to deploy it somewhere so that other people can use it.

2:38:35
I think the easiest option is using the shinyapps.io website,

2:38:40
which is produced by the R studio company.

2:38:45
It has free or paid account tiers

2:38:47
that you could use depending on what you need.

2:38:50
the Delta Science program also does have a paid account,

2:38:53
and they might be willing to host your app if you reach out to them.

2:38:59
Or you can always host an app online on your own server

2:39:03
if you have the time, skill, or money to figure that out.

2:39:09
So some example shiny apps from the Delta.

2:39:13
I have links to a couple of them on the slide here.

2:39:17
I know the Bay-Delta Live website has at least one shiny app.

2:39:21
I think they used to have a couple more, but I'm not sure if they're there anymore.

2:39:25
This hatchery releases app, and then the Delta Science program,

2:39:30
that I mentioned before, they have, a bunch of apps linked there,

2:39:34
and I'm just showing here the top three of them, but there's,

2:39:38
I don't know, like nine, 8 or 9 of them total that you could explore

2:39:41
for some examples.

2:39:45
And then there's a lot of resources for learning

2:39:47
the posit that the company behind RStudio has a lot of good resources there.

2:39:51
The people who develop shiny, this is kind of how I first learned how to use it.

2:39:56
So there's an, a section of articles on shiny.

2:39:59
There's, kind of a tutorial to get started with shiny.

2:40:03
And then they have a gallery of apps that you can explore, which is very nice

2:40:07
to just try to see what you know, what it's possible to be produced.

2:40:12
And then of course, Google is

2:40:14
oh, it's good for finding information.

2:40:20
And that's all that I have.

2:40:23
So thank you.

2:40:25
Thank you Sam.

2:40:26
I hope everyone can have fun kind of exploring some of those shiny apps that Sam shared.

2:40:32
There are a lot of different features

2:40:33
on all of them, and much more complex than what Sam shared.

2:40:37
But just nice to see, what is possible.

2:40:42
So next up we have, Rosie

2:40:45
presenting on R markdown.

2:40:50
All right.

2:40:50
Thanks, folks.

2:40:52
we had a small last minute

2:40:54
change of plans, and now you're expecting an R markdown presentation. But

2:41:00
we were talking through it

2:41:02
and decided that, rather than, starting off

2:41:06
crawling and then, walking, we just go straight to walking.

2:41:10
So, I'm going to be talking about Quarto rather than R markdown.

2:41:15
Don't worry for what we're doing.

2:41:17
It's basically the same.

2:41:18
but there's a lot more functionality in Quarto,

2:41:21
than you get with R markdown as you get more advanced.

2:41:25
It's kind of the the latest, greatest next thing.

2:41:27
and sort of the beginning levels or like I said, pretty identical.

2:41:33
So, this is a lot simpler than a shiny app.

2:41:37
It's just a way to make a reproducible,

2:41:40
report using code because,

2:41:45
I like code.

2:41:46
I can read code.

2:41:49
If you give me a document like this, I'm like, okay, great.

2:41:52
I'll just run it, see what your graphs are,

2:41:54
read a couple of your notes, go from there.

2:41:56
Be great.

2:41:57
But if I give this to my boss, it's not going to go well.

2:42:01
You know, most of the time when we're communicating or communicating with people

2:42:04
who don't want to see the code, or they just want a little bit of the code

2:42:10
and a lot of interpretation and the outputs of the code.

2:42:14
Now, I could make a word document and copy and paste the outputs of my code, graphs

2:42:19
stuff like that into my word document, and type out explanation.

2:42:23
But if I change my data,

2:42:26
I'll have to do all of that copying and pasting and exporting over again.

2:42:31
So automating it is going to be a better way to go.

2:42:36
So with Quarto, you can automatically

2:42:40
build a document that includes text, block,

2:42:44
code and the outputs of code all in one big document.

2:42:48
And then any time you update the data,

2:42:50
you can just hit go and update your entire report.

2:42:53
So for things you have to do frequently, this is awesome.

2:42:57
This is an example of a report that,

2:43:01
I built recently that needs to be updated every day.

2:43:04
And I just have a windows task on my computer.

2:43:07
It reruns my report every single day.

2:43:10
and, I can distribute the results.

2:43:13
So it's pretty slick.

2:43:17
As I mentioned, R markdown was

2:43:19
what everyone used to use for this, you know, five years ago.

2:43:23
and Quarto is built by the same people as R markdown.

2:43:28
It's posit the company that brought you RStudio

2:43:30
and shiny and all the great things in life.

2:43:34
I swear I was not paid to say this, Quarto has a little more functionality.

2:43:38
you can do things like build books and,

2:43:42
more fancy stuff, just a little more easily.

2:43:45
It's sort of a new standard.

2:43:47
It also has more different code languages supported.

2:43:50
You don't actually need to run it in R at all.

2:43:52
If you're a Python user, you can totally use Quarto just from Python

2:43:56
without having to touch R.

2:43:58
but otherwise they're really basically the same thing,

2:44:02
especially if you're just doing, a basic document using R.

2:44:08
And it's

2:44:09
all integrated into RStudio really nicely.

2:44:12
It's set up to make it easy to build one of these.

2:44:15
You just go up to the little new document thing and hit Quarto document.

2:44:21
you fill in a couple of details in the pop up window.

2:44:25
Title, author, what kind of report you're making.

2:44:29
HTML will give you the most functionality.

2:44:31
but if you want a PDF, you can do that.

2:44:33
And word documents are nice.

2:44:35
If you're like writing a paper and going to send this to someone else to edit.

2:44:41
yeah.

2:44:43
And all those details you just entered

2:44:45
there will be entered into the Yaml header at the top of your document.

2:44:50
RStudio puts, like a quick little template together.

2:44:54
It bases the Yaml header off your, details you entered already.

2:45:00
and then it starts

2:45:02
you off with some basic, things that you can edit as you see fit.

2:45:08
In your document,

2:45:09
again, the headers at the top, this is where you put in a lot of things

2:45:14
about what you want your document to do.

2:45:17
Do you want to have it be a HTML or word?

2:45:19
How do you want to edit it?

2:45:21
There's other options like tables of contents,

2:45:24
styles, formats that you can put in here.

2:45:28
This is just for a basic thing.

2:45:32
Then the meat of your document is going to be in code chunks.

2:45:35
RStudio nicely color codes, code chunks versus markdown chunks.

2:45:40
code chunks are indicated by these three backticks

2:45:44
curly brackets, the name of the coding language.

2:45:47
And then at the end of your code chunk, you'll have three more backticks

2:45:51
within that code chunk, it's like you're in an R script.

2:45:54
Everything runs just like an arc script.

2:45:57
The rest, which is sort of white here

2:45:59
rather than kind of beige, is your markdown chunk.

2:46:03
This is where all the text and interpretation in your report goes.

2:46:08
it doesn't do anything.

2:46:10
It's not, code, but it's where you give the explanation,

2:46:15
why are you doing it, what are you doing, etc.

2:46:19
you'll notice that this kind of looks like code, though.

2:46:22
There's like weird symbols in it and hashtags, and it's not very pretty.

2:46:27
but we'd want to report to, you know, maybe be in

2:46:30
Times New Roman and have different font sizes and stuff like that.

2:46:34
Well, you can do all of that using markdown

2:46:37
formatting.

2:46:41
Quarto uses pandoc

2:46:43
markdown standard for markdown formatting, and that's where those hashtags

2:46:48
and asterisks and stuff turn into different formats for text.

2:46:53
there's plenty

2:46:54
of information online for how you do that

2:46:57
Formatting.

2:46:58
You can also use HTML

2:47:00
and get really fancy

2:47:02
or add CSS style sheets. you can.

2:47:04
The sky's the limit

2:47:05
for making really cool looking reports, if you're savvy with that.

2:47:10
But just a couple of things for a basic

2:47:14
quick little report.

2:47:18
Plain text is just plain text.

2:47:20
You need two spaces to start a new paragraph.

2:47:23
Asterisk before and after turn into italics, two asterisk

2:47:27
turns into bold, etc..

2:47:29
the one thing that's really useful is these headers.

2:47:32
One hashtag at the top level, header to the second level header.

2:47:36
And you need a space between the hashtags in the words.

2:47:39
Of course in your code hashtag still indicate comments, but in the markdown

2:47:44
chunks, those are your headers that allow you to build an outline.

2:47:48
Make it easy to navigate your document.

2:47:52
So that's some of the formatting things.

2:47:54
But if you're like me and are tired of googling

2:47:58
how you do bold, R studio has a solution to that.

2:48:03
So right now I've been showing you the source version,

2:48:06
which means the code and the formatting marks.

2:48:10
We can also switch to the visual version of this file,

2:48:13
and you end up feeling like you're in a word processor.

2:48:17
You can easily, say what the header should be.

2:48:22
add bullets, numbers, insert tables, pictures,

2:48:27
links, just like a word processor.

2:48:30
And you can switch back and forth at any time if you want to see

2:48:33
what the code markdown syntax looks like.

2:48:38
So that's a little anatomy, a little markdown.

2:48:41
Now the code which is the most important part

2:48:45
probably your code chunks, you'll notice, have a little green arrow at the top.

2:48:50
This is to run all of the code in the current chunk,

2:48:54
and the little arrow and green line is to run all the chunks above this one.

2:48:59
So set everything up for this chunk.

2:49:02
And a lot of people like using quarto documents

2:49:05
to do all their coding instead of coding in scripts

2:49:08
because, the outputs of this code will drop down

2:49:12
right below the chunk and allow you to go back and refer to them

2:49:16
later instead of having them scroll all the way up in your console.

2:49:22
So I'm going to quickly

2:49:25
go to R studio and show you what that looks like.

2:49:29
sound

2:49:32
Sorry.

2:49:37
So, this is my Quarto demo file.

2:49:40
Please zoom in.

2:49:41
One more click.

2:49:44
just a standard Yaml header.

2:49:49
Information about Quarto.

2:49:50
Here's a code chunk.

2:49:52
It's just going to run one plus one.

2:49:54
I hit the arrow and two goes right there.

2:49:57
I want to do.

2:50:00
Run it again to 11.

2:50:03
I can do something like.

2:50:09
All right.

2:50:10
If I want to just get rid of this

2:50:13
for now, I can hit the X box there and it'll clear it.

2:50:16
You can also add plots.

2:50:18
run.

2:50:20
That gets you the histogram of speed down there.

2:50:26
If you want to add a chunk, you can, actually type Backtick

2:50:30
Backtick, Backtick bracket, R bracket,

2:50:33
or you can go control alt I

2:50:38
or hit this little insert code chunk button at the top to get new code

2:50:42
chunk.

2:50:43
You can also use this to run just the selected lines, current chunks, etc..

2:50:49
So all right, there.

2:50:50
It's got spellcheck.

2:50:52
It's pretty nice.

2:50:54
sounds.

2:50:55
yeah. So that's running code.

2:50:57
And like I said, a lot of people like just having the outputs.

2:51:01
Hang in there.for you to use later.

2:51:05
So that's

2:51:06
if you're working with your code and you're like

2:51:08
messing with things yourself, but when you're getting ready to share it

2:51:11
and turn it into an actual document, you probably want to,

2:51:15
change the way the output to that code looks a little bit,

2:51:19
you might not want all of the code displayed.

2:51:23
For instance, you do if you're,

2:51:27
sharing it with another coder.

2:51:29
But if this is for your boss, maybe you don't want that.

2:51:31
maybe you want the figures to be bigger or smaller or left

2:51:35
aligned or right aligned.

2:51:36
lots of different options.

2:51:40
And to do that, in your chunk, you have the name

2:51:44
of the coding language in the first curly part of the curly bracket.

2:51:48
Then you have the name of the chunk that's optional.

2:51:50
Then you have a bunch of chunk options.

2:51:53
They say how the output is going to be formatted.

2:51:56
And so far as I can

2:51:57
tell, this is the one difference between R markdown and Quarto.

2:52:00
R markdown has chunk options in the curly brackets.

2:52:03
Quarto has them below the curly brackets with hashtag straight line,

2:52:09
but all of the options themselves

2:52:11
are the same important ones to know are echo equals true.

2:52:15
That means do you want to, print the

2:52:20
code as well as the output of the code?

2:52:23
Eval equals true.

2:52:25
Do you want to actually run the code or just print the code?

2:52:28
Include equals true.

2:52:30
This is good for like when you load all your packages.

2:52:32
You don't need that in your final report, but you need

2:52:34
to test from the rest of the code so you can not include it.

2:52:38
then error message and warning.

2:52:40
there's some messages you get every time you plot something.

2:52:45
It says, oh, we removed some missing values.

2:52:48
Not really important to have in your file report.

2:52:50
So you can turn those off.

2:52:53
There's also a

2:52:53
number of figure options for making that look nice.

2:52:57
There's dozens more chunk

2:52:59
options available, and more in various packages.

2:53:04
These are just some of the ones that I use most frequently.

2:53:08
Now, once you have

2:53:10
all of your markdown, set up, all of your code ready to go.

2:53:14
You can, render your quarto document

2:53:18
into HTML or word or whatnot.

2:53:22
this is also called knitting because a lot of these,

2:53:24
tools for rendering are in the knitter package.

2:53:30
so here I have format HTML.

2:53:33
If I wanted a word document, I just put format word, etc..

2:53:36
if you do PDF or word,

2:53:39
you do need to install a latex engine, which is an extra little thing.

2:53:43
It's not too difficult, but it's just something to know.

2:53:47
and I'll just quickly show you how that works.

2:53:55
You have to save it before you can render it.

2:53:57
and then I hit this render button right up there,

2:54:03
and you'll notice that, instead of in the console, you get information

2:54:06
in this background jobs, window that just says what it's doing.

2:54:11
And then you get your, final document,

2:54:18
which is a nice looking,

2:54:23
HTML file

2:54:24
that you can open in your browser with various bits of code.

2:54:28
some plots.

2:54:31
I have some examples in here that's available

2:54:34
on the GitHub of how to make interactive tables.

2:54:37
but this is just the beginning.

2:54:43
that was a very brief demo of a really basic report.

2:54:47
But, if you go to this, quarto document gallery,

2:54:54
there are lots and lots of things you can do.

2:54:56
You can get really slick with, word documents,

2:55:00
advanced layout for HTML and PDFs.

2:55:03
You can make presentations. So

2:55:06
you do your whole

2:55:06
PowerPoint presentation with quarto, including plot.

2:55:10
You can make whole websites

2:55:11
and Perry is going to talk a little bit about this next, as well as web books.

2:55:16
So, I highly recommend kind of scrolling through some of these.

2:55:21
And if you see something you like,

2:55:22
you can look at the code behind it and copy and paste.

2:55:26
and yeah, we can also include,

2:55:30
you know, shiny apps like Sam showed, in your,

2:55:35
things. So

2:55:37
that's all I have.

2:55:39
And I will pass it on to Perry

2:55:44
for our last presentation of the session.

2:55:54
Yeah.

2:55:54
Thanks, Rosie. in.

2:56:01
Media shapes.

2:56:01
Presentation.

2:56:04
Okay,

2:56:05
Great.

2:56:07
All right, so thanks for our presentation.

2:56:08
Really.

2:56:09
I'm going to follow up by talking about something, very similar,

2:56:14
which is publishing reports via book down slash GitHub pages.

2:56:19
I'm Perry, by the way, and I'm an environmental scientist at,

2:56:23
DWR that mostly works with, like, data science type work

2:56:27
and so, what is the goal here?

2:56:30
So all of us work at different agencies and different programs.

2:56:33
But one thing that we all have in common

2:56:35
is that we have to meet reporting requirements.

2:56:38
and this can be very cumbersome to do, like some of us contract out

2:56:42
and have to work with middlemen and such, and it's just like a hassle to deal with.

2:56:46
So what we really want is just a simple, reproducible way

2:56:48
to generate and publish all the reports that we need.

2:56:53
And one solution is using a package called bookdown.

2:56:56
So Bookdown

2:56:57
is an R package that facilitates writing and publishing reports with R markdown.

2:57:01
but there are also partial, partial equivalents,

2:57:05
that work very similarly if you rather use that.

2:57:09
But the advantage of bookdown is that you can create a website to host reports.

2:57:14
You can save these reports as PDFs or word docs or whatever format you want.

2:57:18
Basically.

2:57:19
since it's based on or markdown, you can have figures and tables

2:57:23
and shiny apps, math equations, whatever you would like.

2:57:26
And you do have the ability to customize the format and layout and design.

2:57:31
One thing I want to talk about a little bit more is this first part

2:57:35
where you can create an actual website and how you would go about doing that.

2:57:40
And the nice thing is that GitHub actually

2:57:42
makes this very easy to use for you via GitHub pages.

2:57:46
So GitHub pages turns any repository you have into a website.

2:57:51
and you can easily access this from your repository

2:57:54
by just going to settings and the pages tab.

2:57:57
one great advantage about GitHub pages is that it's free

2:58:00
hosting and easy to use, so no need to pay or anything like that.

2:58:05
in addition, you can choose

2:58:06
a custom domain name, EMP.

2:58:09
my group doesn't personally do that,

2:58:11
but if you would like to, if you have that ability.

2:58:16
All right.

2:58:16
So what I want to do now

2:58:18
is just show you an example of a website that is built via Bookdown.

2:58:22
And then I'll go into the specifics of how you actually go about building it.

2:58:26
And so this is based on, EMP, the environmental modeling program.

2:58:30
current website.

2:58:32
So this is what our landing page looks like.

2:58:34
As you can see, we made it all pretty.

2:58:36
We have a little picture there and everything.

2:58:38
And we have all of our reports, on the side.

2:58:42
And I'm going to focus on the phytoplankton report.

2:58:45
So this is the landing page for it.

2:58:46
we go into the background and things like that.

2:58:49
You can have like bold text, you know, headers, all that fun stuff.

2:58:55
And these are some example sections from the report.

2:58:57
So as you can see we have like math equations and their list

2:59:02
figures with their captions you can put tables etc..

2:59:07
So it's really flexible and the types of things

2:59:10
that you can actually put into these reports.

2:59:14
Okay.

2:59:14
So how do you actually go about publishing, the bookdown

2:59:17
and GitHub pages.

2:59:20
So there are six main steps essentially.

2:59:23
and I'm just going to go through them all, in order.

2:59:26
So the first one is to simply create a GitHub repository.

2:59:30
This repository is the same as any other repository that you would use.

2:59:35
There is no specific way you need to format your files or anything,

2:59:39
but I like to put all the main files into a folder called admin.

2:59:47
The next

2:59:47
thing is to do is to create a Yaml file.

2:59:51
Yaml stands for yet another markup language.

2:59:54
You don't really need the specifics of it, besides the fact that it basically tells

2:59:58
the website how to render your file.

3:00:01
And so the main Yaml is broken up into two parts.

3:00:04
You have like metadata

3:00:05
that's just like the file names output directory, things like that.

3:00:09
And then you have the actual markdown files that contains the report.

3:00:12
So every single one of these lines from 15 onward,

3:00:16
contains one of my reports.

3:00:21
The next thing you can do is

3:00:22
you can specify additional Yaml via an RMD file.

3:00:26
you don't have to necessarily do it this way,

3:00:28
but I find it the one with the most customization options.

3:00:33
And, just

3:00:34
to put that in, context, that's what this,

3:00:38
RMD file is, that is currently highlighted in the red box.

3:00:43
So that's what this file looks like.

3:00:45
It's mostly just a bunch of metadata information.

3:00:48
but I just wanna highlight two things very quick.

3:00:51
One is this line here with the under header,

3:00:55
this links to, a Google Analytics file.

3:00:58
So if you want to track who's going on

3:01:00
your website, how many hits you're getting on it and things like that,

3:01:03
you are able to integrate that within your report.

3:01:08
The other one I want to highlight is this part, the CSS.

3:01:11
So CSS, designed your website.

3:01:14
Essentially it specifies the layout and everything.

3:01:18
this is what CSS it looks like.

3:01:20
It's a whole different language and very complicated.

3:01:22
R does give you some default templates

3:01:24
you can use if you like what R report uses.

3:01:28
You can also just use it as a template, so feel free to do that.

3:01:35
The next thing you want to do after you specify the backbone of your website,

3:01:37
is to actually create the reports as markdown files.

3:01:41
And again, just to put it in context, that's what all these little lines

3:01:44
15 onward, are referencing.

3:01:49
So this is an example of it

3:01:50
in the visual mode of an R markdown file.

3:01:53
and just to highlight some specific parts of this.

3:01:56
So one thing to do is you can just write text normally

3:01:59
you can build it, make headers, all that fun stuff.

3:02:03
You can also add inline R functions if you want to do it that way

3:02:08
you can add your figures and tables just as normal.

3:02:10
r markdown code chunks.

3:02:14
And so yeah.

3:02:15
So basically anything you can do in markdown

3:02:17
you can just do here and it'll render as part of the website.

3:02:23
Now that you're all done and you made all your files,

3:02:25
now you just need to run the main R file to actually render the reports.

3:02:28
And this would also publish it to GitHub pages.

3:02:32
So, this is what R file looks like.

3:02:34
You don't have to worry about the top part,

3:02:35
but this one function render book is basically all you need to do.

3:02:40
As you can see to the inputs are those Yaml files I talked about earlier.

3:02:43
Output format just tells it what format you would like, so get book

3:02:47
will render it for HTML pages for website, but you can also do PDF for word docs.

3:02:52
whatever you really want.

3:02:54
and the last one I want to highlight is this part,

3:02:57
that output directory, that is set to dot docs by default.

3:03:02
to put that into context, that's what

3:03:04
this folder is here in the main repository.

3:03:08
And that is where all of the files actually build

3:03:11
he website will get saved to once you render and you specify within the,

3:03:16
GitHub repository setting that that's the page you want to use

3:03:19
under this little branch, heading right here.

3:03:24
And then once that's finished,

3:03:25
all you need to do is check that your deployment was successful.

3:03:29
So if you actually

3:03:29
go back to any repository, you'll see down in the right hand corner

3:03:33
you have a little section called deployments.

3:03:35
And if you click on that you'll get a page like this.

3:03:38
And if the top part is a little green checkmark, the last deployed

3:03:43
gives you the name of the website and you're deployment was successful.

3:03:47
Okay.

3:03:48
So that's that's all you really need to do is pretty simple.

3:03:53
And. Yeah.

3:03:53
And then just, some additional resources.

3:03:56
This is the bookdown reference site.

3:03:58
Here is the CSS template or EMP website.

3:04:02
And if you want to do this in Quarto as opposed to

3:04:04
R markdown, it's a little different but overall similar.

3:04:08
And you can find those references here.

3:04:11
really quick just to show you

3:04:13
this is what the actual website itself looks like.

3:04:17
So you can kind of see what it looks like when it's all put together.

3:04:21
Okay.

3:04:22
So yeah.

3:04:35
Thanks, Perry

3:04:35
Can you just put that link, for your site in the chat?

3:04:40
I think the link

3:04:41
that I have is broken or something, because it keeps coming up with an error.

3:04:44
So I just want to update that

3:04:47
in our references.

3:04:54
And I

3:04:55
think we'll have to do a quarto book making training at some point.

3:05:01
Now things are changing.

3:05:05
So very, very similar.

3:05:07
And if you, if you,

3:05:11
yeah, if you understand the R markdown version,

3:05:13
it will take you three seconds to pick up the quarto version.

3:05:17
Okay. Great.

3:05:19
Yeah.

3:05:20
The basic steps are still the same.

3:05:22
Either one.

3:05:24
Okay. Yeah.

3:05:25
I'm in the process of building a book, so it's maybe a good time

3:05:29
to be learning about this and transfer over.

3:05:34
So, Susannah,

3:05:36
go ahead and start sharing.

3:05:48
Okay.

3:05:48
So you see. Okay.

3:05:51
Yep. Looks good.

3:05:52
Okay, great.

3:05:53
So, hello, I'm Susannah Iltis, the web computing specialist for Columbia Basin

3:05:58
Research at the University of Washington and the group behind SacPas.

3:06:03
Today, I'll be presenting an overview of SacPas and data visualizations.

3:06:07
And I hope that you'll learn

3:06:08
something new in terms of our process and specific tools available.

3:06:12
I'll start with a bit on the background of SacPas, and then our

3:06:15
considerations for data and visualizations before diving into SacPas websites.

3:06:20
Specific tools and examples.

3:06:25
I want to take a few moments to introduce Columbia Basin Research to you.

3:06:29
The CBR website started in 1994 with Columbia River

3:06:33
DART Data access in real time and other products.

3:06:36
DART is a secondary data repository with centralized

3:06:39
and integrated fish data and environmental conditions.

3:06:42
Related to river, estuary, ocean and climate.

3:06:46
We mirror data from federal, tribal, state, governmental agencies

3:06:49
and other entities to aggregate and integrate data

3:06:52
into a centralized location accessible to the public and real time.

3:06:57
And we started a similar effort with USBR for the Central Valley in 2006.

3:07:02
And that SacPas, our database manager, web specialist

3:07:06
and CBR researchers prepare the data for analysis and online tools.

3:07:11
And these are created with as much foresight as possible

3:07:14
and in collaboration with data owners and stakeholders.

3:07:17
Over the decades,

3:07:18
we've developed many different tools, and we continue to update and evolve

3:07:22
with a lens on human centered design and how best to communicate science.

3:07:28
And I want to mention that updates to the SacPas

3:07:30
website are in progress, with implementation plan for this fall.

3:07:36
So on the website we offer tools and services

3:07:39
in multiple sections including data, queries and alerts, workgroups and teams,

3:07:43
fish modeling tools. In the Fish model and tool sections,

3:07:47
we have models and exploratory shiny apps developed by CBR,

3:07:51
and there are also third party models.

3:07:53
The Loss and Salvage Predictor and Delta Star models are created by others

3:07:58
and then deployed on SacPas to make them publicly available with real time data

3:08:02
inputs, as well as to automate in-season model

3:08:05
runs and results for use by the salmon monitoring team.

3:08:10
Today, I'll be focusing on the data

3:08:12
queries and alerts in the work groups and team sections of the website.

3:08:17
Next, I'd like to go over considerations that we take with regards

3:08:21
to data and visualizations and show an example of our process to product.

3:08:26
All the data visualizations on SacPas are custom,

3:08:29
from simple line plots to heat maps and others.

3:08:32
When developing visualizations to communicate information,

3:08:36
we first considered the purpose of the tool and the types of users

3:08:40
for management teams that can include specific and relevant metrics

3:08:44
and integrated data sets, and for research scientists, access to data downloads

3:08:49
and for the public, equal access to data visualizations and modeling tools.

3:08:55
In temporal scale for pre-season and in-season management is another facet

3:08:59
of data and purpose, whether it is real time, historical or forecast.

3:09:04
And for spatial scale.

3:09:05
We started with the Sacramento River and then expanded into the San

3:09:09
Joaquin and the Delta.

3:09:12
For the design of

3:09:13
visualizations and data products, we focus on relevancy,

3:09:17
which is enhanced through collaboration and user feedback.

3:09:21
Usability as providing meaningful summaries and designing clear visualizations.

3:09:26
Accessibility.

3:09:27
Creating for an audience with diverse abilities in mind

3:09:31
and structuring for quick, intuitive online interactions

3:09:34
and transparency, including data attribution and reference methods.

3:09:40
As an introduction to how we work,

3:09:42
I will cover one real world example of process to product.

3:09:46
The product is this figure which shows daily Caswell RST

3:09:49
total juvenile catch and refined flow on the bottom

3:09:52
and the distribution of catch by size and life stage on the top,

3:09:56
with trapping dates indicated by the plus symbol in between.

3:09:59
The process started with the Stanislaus Watershed team meeting with the SacPas

3:10:04
to discuss initial team needs, and was followed by additional discussions

3:10:08
as one result.

3:10:10
The SacPas team reached out to Pacific State's Marine Fisheries Commission

3:10:14
Caswell RST data contact.

3:10:16
We inquired whether it be possible to provide the data publicly

3:10:19
during the season, and if there was any publishing requirements or concerns.

3:10:23
The SacPas team reached out to Cal Fish, where

3:10:26
the RST data sets from the region are made publicly available

3:10:30
with a request to coordinate with Caswell RST

3:10:33
data contact to establish a data upload process.

3:10:37
Once the data set was made publicly available, this SacPas

3:10:40
database manager designed table structures for the data, developed

3:10:44
procedures for retrieval, processing and loading into the local database,

3:10:48
verified with the primary source, and implemented daily updates.

3:10:53
And then finally the SacPas web computing specialist.

3:10:56
Okay, me designed the visualization based

3:10:59
on existing figures produced by PFMC that was distributed via email

3:11:04
and distributed query and visualization code, and implemented

3:11:09
daily updates on the website, producing this figure to make it public.

3:11:15
Now I'll dive into tools and visualizations from our data

3:11:18
queries and alerts and our work groups and team sections of the website.

3:11:22
To start, this is a snapshot of data available on on the SacPas website.

3:11:27
For environmental data, we offer a selective list

3:11:30
of CDEC and USGS locations and data.

3:11:33
Please let us know if there's any that you would like us to add.

3:11:37
We also have, Delta status and operations from DWR. For fish data,

3:11:42
we have juvenile monitoring for RST, trawls, beach seines in Red bluff

3:11:47
and juvenile.

3:11:48
Sorry.

3:11:48
We have salvage and loss, adult escapement,

3:11:51
Red and carcass surveys and delta smelt.

3:11:54
And the fish data is retrieved from the data owners

3:11:57
Cal fish and the EDI data portal.

3:12:03
The data queries and

3:12:03
alerts section is organized into six topic areas.

3:12:07
For fish,

3:12:07
we have juvenile monitoring and sampling, juvenile salvage and loss,

3:12:11
adult escapement, and for water we have temperature thresholds,

3:12:15
river conditions and exposure index.

3:12:18
Today I'll be

3:12:20
covering products from three of the areas the juvenile monitoring and sampling,

3:12:24
juvenile salvage and loss and river conditions.

3:12:28
Before talking about individual tools, I'd like to cover

3:12:31
common elements shared by many of the SacPas data query interfaces.

3:12:35
That's query title or type.

3:12:37
Primary data source attributions.

3:12:39
Available topic area queries.

3:12:42
Available output formats for viewing online and downloading selections.

3:12:46
Customized to the data in the query and the Submit Query button.

3:12:50
And then to the right of the submit button

3:12:53
a check box to generate a reusable query URL for automated retrieval

3:12:58
based on current selections, and as you heard this morning,

3:13:01
you can see here a snippet of R code in order to implement that

3:13:05
and then query and query notes and resources relevant to the query.

3:13:12
in the Juvenile Monitoring and sampling topic area,

3:13:14
we provide access to current and historical data collected

3:13:17
by multiple agencies at multiple locations.

3:13:21
Visualizations and data reports are customized for multiple purposes,

3:13:25
and the next four slides I will touch on the connection

3:13:28
between three of the data products the cohort juvenile monitoring,

3:13:32
the migration timing and conditions, and current catch.

3:13:37
The Cohort Juvenile Monitoring product provides a single point access

3:13:41
to rotary screw traps, beach seines, trawls and the Red Bluff Diversion Dam

3:13:45
data collected by multiple agencies.

3:13:47
The tool is designed to view a single brood year cohort at multiple locations.

3:13:52
SacPas creates the brood year cohorts based on Red Bluff Diversion,

3:13:57
diversion dam and length at date

3:13:58
model run assignment for tracking the migration downstream.

3:14:02
The figure shows fish presence, cumulative catch and timing at multiple locations.

3:14:09
The Migration Timing

3:14:10
Conditions product uses the same brood year cohorts to calculate

3:14:15
cumulative percent of catch or passage estimates over the cohort date period.

3:14:21
To compare estimates of migration timing across historical years.

3:14:25
The visualization is a combination of two plots.

3:14:28
The main plot on the left shows for each brood year,

3:14:31
the cumulative percent represented by the horizontal bars,

3:14:34
and a vertical bar on the 50% day

3:14:37
and oh for the cohort date range period.

3:14:40
And the plot on the right shows the total catch of

3:14:42
Passage estimates for the period.

3:14:47
The current catch visualization combines the historical catch

3:14:50
and timing with current year catch at multiple locations for tracking

3:14:54
the downstream migration of the cohort from Red bluff, to Knights

3:14:58
landing to Sacramento trawls to Chipps Island trawls,

3:15:03
and then providing a simple prediction for the current brood year

3:15:06
based on the average estimated historical timing and range.

3:15:10
Zooming in, we can better see the brood

3:15:13
Year 2023 prediction on April 9th for winter

3:15:16
run Chinook at Chipps Island in numbers and chart. The color and pattern

3:15:20
are used to differentiate historical and current data from prediction.

3:15:27
And the juvenile salvage and loss topic area,

3:15:29
we offer data visualizations and summary and detail reports

3:15:33
for the CDFW

3:15:36
salvage database, combined with water data

3:15:39
and Delta loss with exports and OMR Index.

3:15:43
We have historical salvage timing by species, run and clip status.

3:15:47
We have salvage and loss detail report

3:15:50
with length and date, CWT and DNA races and the Delta salvage

3:15:56
with flows and export.

3:15:59
The Chinook Delta salvage data

3:16:01
visualization shows the distribution of salvage by size and date

3:16:05
for known and unknown origin, hatchery and unclipped Chinook, as well as the DWR

3:16:11
length at date Delta model, the daily

3:16:14
DCC gate status and then multiple flows.

3:16:18
As you may have noticed, titles can be detailed because the online plots

3:16:22
can be considered as standalone products, so all the information that are contained

3:16:27
in figure captions are specified in various places on the plot,

3:16:32
including the data source and timestamp

3:16:34
of when the query was run and the product was produced.

3:16:38
More detailed information is available on the associated web page.

3:16:42
We strive to be as transparent as possible

3:16:45
and cite data sources and methods where appropriate.

3:16:50
In the River Conditions

3:16:51
topic area, we offer multiple tools to access

3:16:54
a selective list of locations and data from from CDEC and USGS.

3:16:59
Our most flexible and configurable query is the River conditions

3:17:03
Graph and Text query.

3:17:05
It contains multiple output formats for viewing visualizations

3:17:08
and data tables online, and for downloading.

3:17:12
There are multiple selections allow for each of the menus

3:17:15
for year, location, and data parameter, and there are additional

3:17:19
selection settings and options for further customization, including

3:17:24
date range, ten year averages, and graph options.

3:17:30
A companion product

3:17:31
is the map interface for the River conditions query,

3:17:35
with the Selection filter option turned on its default state.

3:17:39
The tool provides SacPas data, inventory exploration responsibility

3:17:43
to user selections for location, data type, and year.

3:17:48
The map interface also allows for querying with the same output types

3:17:51
as found on the main query, and the next three slides,

3:17:55
I will cover an interaction with the interface to answer the question

3:17:58
at which locations and for which year is the data type.

3:18:02
Tidally filtered flow available?

3:18:05
Before we start, I'd like to point out that with no selections made,

3:18:08
there's 122 possible locations.

3:18:12
So first we select the tidally filtered flow from the data type menu.

3:18:17
After the selection, we see great reduction of sites

3:18:20
available on the map and the location menu.

3:18:23
The year menu is also responsive to the selection.

3:18:28
Next, we

3:18:28
select our locations of interest from the available options

3:18:31
in the location menu or from sites on the map.

3:18:34
I have selected both the USGS sites

3:18:37
at Freeport and Rio Vista.

3:18:40
The submit button at the bottom of the control

3:18:42
panel is still gray, which indicates that all elements

3:18:46
required for a query have not yet been selected.

3:18:50
Returning to our question and the second part what years are available?

3:18:54
We can now see that there are ten years in the year Menu.

3:18:57
Based on our selections for data type and location.

3:19:01
After selecting one year in the menu, the submit button is no longer gray.

3:19:05
It's now active and clickable.

3:19:07
Submitting a query request is now possible

3:19:10
with the selection filter option turned on.

3:19:12
Receiving a no data available result should not be possible.

3:19:16
And I'll note that the zoom level is controlled manually by the user.

3:19:22
So whether

3:19:23
using the main query page or the map interface, submitting a request produces

3:19:27
the same visualization data table or download and in our example here

3:19:32
we have the USGS at Freeport and Rio Vista for tidally filtered flow.

3:19:37
And the years 2022, 2023 and 2024.

3:19:44
I'd also like to highlight the All Years

3:19:46
River graph, another visualization tool in the river conditions area.

3:19:51
It is designed for visual comparison of conditions

3:19:54
for single location and data parameter.

3:19:57
In this example we are viewing water temperature

3:20:00
at Rio Vista Bridge for all water years to the current day.

3:20:04
On the left we have months from October through September,

3:20:07
and on the right the color legend and scale.

3:20:11
Looking horizontally across water years, we can quickly see certain years having

3:20:15
among the highest or most temperatures during certain periods of the water year.

3:20:22
One of the customization options to view the data

3:20:24
in potentially more meaningful way is to enter a threshold value.

3:20:29
In this example, the water temperature at Rio Vista Bridge

3:20:32
the same as the previous slide.

3:20:34
The threshold is set to 56°F, and the period of interest

3:20:38
is set to December through May.

3:20:43
And this is another example from the tool where the color palette is

3:20:46
customized to the data type, in this case electrical conductivity

3:20:51
at Emmaton.

3:20:55
So these two

3:20:57
river condition visualization tools allow the user

3:21:00
to examine the data in different ways and answer different questions.

3:21:03
In this example, the data set is the same.

3:21:06
Flow at USGS near Vernalis restricted to January through May.

3:21:11
The figure on the left shows the current year, with the ten year

3:21:14
average produced by the river conditions graph and text tool,

3:21:18
and the figure on the right shows the full historical data set

3:21:22
for comparison across year, month and value produced by all years

3:21:26
river graph tool.

3:21:28
Accompanying both visualizations is the option to

3:21:31
to have a download file of the data.

3:21:36
I'd like to say so.

3:21:37
Customized tools for requests enhance

3:21:39
the relevancy and usefulness of past products to the region.

3:21:43
The weir overtopping Alert tool

3:21:46
is the original, customized tool request that is comprised of a web page,

3:21:50
including National Weather Service forecasts of River stage,

3:21:54
the last ten days, and the water year to date for four for four weirs.

3:21:59
In addition, there's a subscription alert service

3:22:02
for daily email when any weir has crested, and the crested days

3:22:06
here are shown highlighted in orange.

3:22:11
In the work Groups

3:22:12
and team section, we provide customized data reports, visualizations

3:22:16
and model output for the Salmon Monitoring Team, Smelt monitoring team, Stanislaus

3:22:20
Watershed team and products for the San Joaquin

3:22:23
River Restoration Program are currently in development.

3:22:27
When requests are made, we

3:22:29
aim to provide process and product that are automated, repeatable,

3:22:34
public, consistent and current

3:22:36
to create content that is relevant to the team, supports

3:22:39
the team decision processes and reduces the burden of repetitive manual creation

3:22:44
where possible.

3:22:46
All products and services are designed, developed and refined in collaboration

3:22:50
with team members and liaisons.

3:22:54
Customized products for the Salmon monitoring team cover early,

3:22:57
middle and late management season reporting from October through June.

3:23:02
Shown here is the water year 2024 natural steelhead

3:23:05
cumulative loss to date with 2 date period

3:23:09
Single year loss thresholds

3:23:11
focusing on the populations for Sacramento

3:23:14
and San Joaquin and the historical losses

3:23:17
providing a prediction and bound for future

3:23:20
dates.

3:23:24
Customized products for the smelt monitoring team include

3:23:27
reporting of Enhanced Delta Smelt monitoring, Chipps

3:23:31
Island trawls catch updated daily, as well

3:23:34
as many customized visualizations of hydrologic conditions.

3:23:38
Shown here as part of the EDSM current water year

3:23:41
detail table, including the data source and data quality.

3:23:45
We, Recent meetings and collaborations with U.S.

3:23:49
Fish and Wildlife Service Lodi Office improve the quality of the data

3:23:53
on SacPas and the timeliness of reporting.

3:23:57
The current and historical Delta Smelt Surveys

3:24:00
query was developed in collaboration with the Smelt Monitoring team

3:24:04
and is currently linked to on the team page,

3:24:08
and has multiple options for further refinement.

3:24:12
The query offers detailed reporting of the Delta smelt

3:24:15
for current and historical surveys and studies,

3:24:18
and the available surveys and dates that we have in the SacPas

3:24:22
data are shown here.

3:24:24
We have a couple of that have been the Fall Midwater Trawl

3:24:28
And the Townet Survey

3:24:32
that we need to update for the previous year.

3:24:37
Products for the Stanislaus Water

3:24:38
shed team include reporting on current river conditions for multiple locations

3:24:43
and parameters, and multiple customized, customized visualizations

3:24:47
of water temperature for both current and historical conditions.

3:24:52
The moving 60 day windows before and after the current calendar date figure

3:24:56
provides a visual prediction of near-term conditions for water temperature,

3:25:01
with minimum and maximum date period mean identified by the dotted

3:25:05
and dashed lines, respectively, and the current year as the solid black.

3:25:12
In conclusion, I hope through this presentation and examples of our work

3:25:16
that I have demonstrated our focus on communicating information intentionally

3:25:20
and meaningfully through visualizations and other tools.

3:25:24
I shared a number of tools under the data

3:25:26
queries and alerts, and the work groups and team section of the website.

3:25:30
Many of these are real time tools and context of historical and sometimes

3:25:34
forecasted, with data summaries and metrics to support decisions.

3:25:40
So thank you for your time today.

3:25:42
We hope to hear from you.

3:25:44
Any errors or bugs you happen to see?

3:25:46
Please let us know and let us know!

3:25:47
If you'd like to see refinement on tools.

3:25:49
We look forward to opportunities

3:25:51
to improve SacPas and for new collaborations.

3:25:54
We thank USBR for funding our SacPas work.

3:25:57
And we thank Cat,

3:25:58
Elissa and Josh and many more for their guidance, support and working with us.

3:26:07
Thank you Susannah.

3:26:09
Great overview.

3:26:10
There we go. I.

3:26:15
Okay, so

3:26:15
our next portal we're going to learn about is Bay Delta Live.

3:26:20
Heard a little bit about it so far.

3:26:22
So, see a little demo from Amye.

3:26:38
Just making sure my mic is good.

3:26:41
Yeah, yeah.

3:26:42
Ready?

3:26:43
Game.

3:26:43
Great.

3:26:50
All right.

3:26:51
Good afternoon everyone.

3:26:52
My name is Amye Osti.

3:26:54
I am a program manager under, the Bay Delta live,

3:26:59
platform and have been working within this industry

3:27:03
and on this particular project for over ten years.

3:27:07
So it's been a long time in development and it's had many iterations

3:27:12
over the years.

3:27:13
And so what we're going to talk about today is.

3:27:19
Oh there we go.

3:27:23
just a little bit brief overview of our user community.

3:27:26
background on what BDL is.

3:27:30
Our partners and investments that are being made in the platform moving forward.

3:27:35
what is BDO doing now and what we plan to do in the future?

3:27:40
And, just show a few of our most current

3:27:43
projects, in the, in the making.

3:27:46
So BDL for being such a unique and a niche market,

3:27:51
we actually have a really robust user community.

3:27:54
Over the past few years, we've had about 300,000 unique users

3:27:59
with 690,000 requested files.

3:28:03
And we do filter out as much and

3:28:06
get into the unique IP addresses for these statistics and then count

3:28:10
all of the various data download, whether it's a data package,

3:28:14
image, PDF, or, data request.

3:28:21
So BDL has been around for the last ten years, like I just said.

3:28:25
And over the years we've had investment in this particular platform

3:28:29
for from all different types of agencies and,

3:28:33
community and NGOs.

3:28:37
all of these investments have resulted in various platforms

3:28:42
that are similar to BDL, but BDL's the engine behind

3:28:45
these various platform's. From the Sacramento River watershed to various

3:28:50
work up in the upper watershed through RCDs and forest health.

3:28:55
And so every time there is an investment in, BDL, all that information,

3:29:00
whether it's a tool, a data set is shared back to the community.

3:29:05
So if we do, something for adding additional

3:29:09
GIS or data sets like the water data library,

3:29:13
our users and platforms throughout the watershed will be able to use it

3:29:18
like Battle Creek or, salinity management or any of the aquatic veg mapping.

3:29:24
All get to use that in their various platforms.

3:29:30
So what is BDL?

3:29:32
BDL is an open data access and visualization platform.

3:29:36
We are an API consumer, so a data federation,

3:29:39
we've been working over the years to actually work with all of the

3:29:44
very robust and successful open data programs to aggregate that data

3:29:48
from hundreds, literally hundreds of

3:29:51
repositories,

3:29:54
individual data sets, the open data

3:29:57
platforms like see CNRA Data.gov,

3:30:01
and NWIS, CDEC, NOAA bring those into one

3:30:05
UI so that you can start to look spatially at these data together.

3:30:10
we tried to create a common operating picture

3:30:13
with current conditions of the delta.

3:30:16
We have done that also

3:30:17
in the upper watersheds for forest health and watershed management.

3:30:21
These current conditions can be visualizations, including river

3:30:26
conditions, weather, operating and,

3:30:30
have them all laid out in a dashboard.

3:30:33
The web application provides,

3:30:35
federated data through custom templates, dashboards and maps.

3:30:39
So we've built a lot of tools on top of these data

3:30:42
so that you can use them, discover, visualize and report.

3:30:48
Much of the

3:30:48
work that we, do is from sensor network observations.

3:30:52
So using the various monitoring programs in one location,

3:30:58
everything we try to, present is map based or spatial,

3:31:02
so that you can do, the visuals through the map

3:31:06
and look at a particular location and then analytics on top of that.

3:31:10
And then the platform itself provides a collaborative environment

3:31:15
for special projects like Forest Health, Battle Creek Fisheries, CSAMP,

3:31:19
Islands, remote sensing, aquatic veg mapping.

3:31:23
So if you have a very specific topic that you're looking to bring

3:31:27
these disparate data together and create a story or an environment

3:31:31
to actually look at specific data, we have the, tools for you

3:31:35
to actually create that channel or individual picture for your teams.

3:31:42
So most people are

3:31:44
familiar with BDL for its data reporting dashboards.

3:31:47
They present disparate data in one UI.

3:31:50
So we just try to bring together as much information on a particular topic

3:31:55
and present it in a nice, clean, interface so that you can get a daily update of

3:32:01
what's happening, like, common operating picture, conditions,

3:32:07
easy to report out who's using it, mainly managers.

3:32:11
BDL has an application for both Android

3:32:14
and iPhone, and at the end there's a QR code.

3:32:18
So you can actually snap it and then, download the app.

3:32:22
Many of these dashboards are available on the app itself

3:32:27
and are used for people just doing a quick check in of what particular,

3:32:31
things are happening with operations, hydrology, exports.

3:32:38
Just another snapshot of a data dashboard.

3:32:41
This one's displaying environmental indicators for fish

3:32:45
migrations and Old Middle River operations,

3:32:49
etc..

3:32:53
you might also be familiar with the Constituent Tracker that's on Delta Live.

3:32:57
And this is a tool used for drought monitoring and,

3:33:00
monitoring for the first flush.

3:33:03
It's basically a data assimilation model using CDEC and USGS.

3:33:09
And then hopefully by the end of the year, more information from the water

3:33:13
data library to spatially look at constituents in the delta.

3:33:18
So you can look at an animated spatial map in real time

3:33:22
near real time, 15 minute intervals,

3:33:25
or at the constant point in tide of turbidity salinity,

3:33:29
so that you can actually see the changes in conditions

3:33:32
with the tide happening, in real time.

3:33:36
So this particular spatial map and all of its corresponding modeling

3:33:41
and graphing tools are used for drought monitoring trawl managers

3:33:46
to see what the actual conditions are at a particular location,

3:33:51
the science community and water operations.

3:33:57
BDL also has a

3:33:59
NASA.baydelta live.com, which is a subdomain,

3:34:03
for all of the remote sensing products that we've been aggregate

3:34:07
that are relevant to the to the Bay delta.

3:34:10
So you can get the most recent and up to date

3:34:13
spatial maps, time series and visualize those over time.

3:34:19
And there are dashboards specific

3:34:21
to algorithms that are looking at turbidity, chlorophyl

3:34:24
and the various applications for, remote sensing in the science community.

3:34:30
You can access each of these data sets.

3:34:33
You can download the images yourself and use them on your desktop.

3:34:38
And, we're adding to this on a regular basis,

3:34:42
including eco stress and additional aquatic vegetation

3:34:45
mapping.

3:34:50
The the platform also has over the years,

3:34:53
has, acquired a robust amount of geospatial data.

3:34:58
So in addition to the Landsat and all of the remote sensing data,

3:35:03
I just spoke about, we've pulled and pull the services

3:35:07
from CNRA Open Data.gov,

3:35:11
all of the GIS that we found relevant or has been requested by various workgroups

3:35:16
so that you can load the data in the map

3:35:19
UI and then integrate it with other,

3:35:23
discrete monitoring, continuous monitoring information.

3:35:26
So we're trying to make sure that you can build a really clear picture

3:35:32
using GIS based data, and then adding all of the monitoring information

3:35:38
On top of that. You can search the data catalogs.

3:35:41
Everything we have in the, geospatial resources

3:35:45
is available for download or used as a service, in your ArcMap.

3:35:55
in addition to just doing basic

3:35:57
geospatial discovery, you we have analyzed data tools.

3:36:01
So on the on the left-hand side here of in the map section of Bay Delta Live,

3:36:07
you'll have a map navigation, very common layout of how to access GIS data.

3:36:12
So you can search for data loaded onto your map.

3:36:16
And then we have these analyze tools.

3:36:18
The analyze tools are very specific to a particular analysis

3:36:23
that's been requested or is, of interest to the community.

3:36:28
So you can analyze islands you can look at for

3:36:31
project information, you can do water quality and abundance.

3:36:35
And basically you just draw, either you can click on a particular feature

3:36:41
or you can draw around a set of features and get some information

3:36:45
back in a report, which can be downloaded and added to your document library.

3:36:53
This is just another example is that once you've loaded

3:36:56
some particular information, you have not only so you have your

3:37:02
project polygons, you can get the data in a table format

3:37:06
which is searchable, download and find any information and zoom

3:37:11
to so if the feature data has robust metadata and or attributes.

3:37:17
They're all searchable, and you can get that information

3:37:20
easily using the attribute table.

3:37:23
features.

3:37:26
So, 2024 is a big year for Bay Delta Live.

3:37:30
We have,

3:37:33
started on some pretty large projects that will help bring,

3:37:38
additional information that's been requested over the years.

3:37:41
And so one is the explore dot Bay Delta Live.com.

3:37:45
This we hope to have everything up and running for the Water Data Consortium

3:37:51
conference that's in the late summer

3:37:55
and explore will provide

3:37:57
separate from all of the dashboards that you find on Bay Delta Live.

3:38:01
Right now, you can go to just explore and it will be very data specific.

3:38:06
You'll be able to search any of the data repositories that we federated

3:38:10
and then use the tools on top of it, whether it's the

3:38:14
the graphing tools or other visualization tools.

3:38:17
And make sure that you can actually use that information.

3:38:21
Download it, and then save for your own purposes,

3:38:27
we'll be operationalizing the water data library

3:38:29
so that it can be used with other data on the site.

3:38:32
So including CDEC you can run comparisons between

3:38:36
the real time CDEC data and the QAQC.

3:38:39
We're looking to operationalize ECOSTRESS

3:38:43
remote sensing with the, remote sensing platform,

3:38:48
adding additional aquatic vegetation mapping.

3:38:51
We are working with the Lodi office on EDSM

3:38:56
DJFMP weekly reports, and daily reporting dashboards,

3:39:00
which I'll show you a little bit at the end of the presentation here,

3:39:03
integrate OpenET models so that you can actually run those models

3:39:08
with existing data on the site, adding some LiDAR.

3:39:12
This these are projects specific.

3:39:14
And then these all go to the upper watershed.

3:39:17
So we're working with the Sierra Nevada Conservancy on the upper watershed to

3:39:22
do a big lidar assessment that will help people understand watershed conditions

3:39:28
when, working for water quality, water operations and water supply.

3:39:36
So explore dot Bay Delt Live,

3:39:38
This is just give you a quick snapshot of what that will look like.

3:39:42
You'll be able to explore any of the data catalog.

3:39:46
So CDEC, NWIS, NOAA, US

3:39:49
Fish and Wildlife. We have, close to 30 catalogs up right now.

3:39:55
And we'll be adding, the water data library and some others

3:39:59
that are probably more relevant to the work that you're doing now.

3:40:03
And basically everything is run on a timeline so that you can click

3:40:07
your start and end dates, and it loads all the station information.

3:40:11
And then from there you can start to use the graphing tools

3:40:14
and other tools that are on the left hand side

3:40:17
here.

3:40:20
So for example, if you were to go into USGS,

3:40:24
NWIS, which is the National Water Information System.

3:40:27
You can choose your, your region or your sensor, start to map it

3:40:33
and then use the timeline to actually hone in

3:40:36
on the time series information that you would like to to load on the map.

3:40:42
Once that information is loaded at that,

3:40:45
you can start to select based on all of the point data here on the map.

3:40:49
So if I pick a time frame from, September to April 2023,

3:40:56
and then it will chart each of the stations that I selected

3:41:00
on the right hand side and you can continue to add that information.

3:41:04
So right now we're looking at tidally filtered flow, EC at Old River,

3:41:09
tidally filtered data at Mokelumne and turbidity.

3:41:15
This data is also made available in a way in an array

3:41:18
so that you can download it and import it into R or a,

3:41:22
your other desktop applications where you're doing additional work.

3:41:27
You can save these maps and they end up in your map catalog.

3:41:31
And then go back and load those maps and it will bring it to current date.

3:41:36
So you have options of how you can save this data.

3:41:40
And then you'll always go back to that saved map and say,

3:41:42
oh, I need that data set again and do your download.

3:41:44
And this is your what you're looking at here on the left hand side.

3:41:50
And then

3:41:51
obviously you just don't look at one data set.

3:41:54
You have the ability to compare multiple data sets with fisheries.

3:41:58
So, what we're looking at right here is we're loading CDEC, NWIS,

3:42:03
and, NOAA information with the EDSM trawls and DJFMP trawls.

3:42:09
So it gives you the opportunity to actually look at conditions

3:42:12
during a particular monitoring program

3:42:15
so that you get the real visual and you can do it in a time series.

3:42:20
So you're looking and using your, timeline to not only see

3:42:24
how the monitoring happened over time, but what were the conditions, and then

3:42:28
potentially loading up any other GIS with the catalog on the left.

3:42:33
So if you want to know what the AG land use was or

3:42:38
what's happening at a particular island, or load the aquatic veg map

3:42:43
so you can see what the aquatic vegetation, the turbidity,

3:42:48
remote sensing products that are developed by JPL and others.

3:42:53
You can start to see how you start to piece

3:42:55
all of this together and get a good visual of the data.

3:42:59
And so you're not always looking at charts and graphs and zeros and ones.

3:43:03
You might be able to start to look for connections and really explore

3:43:07
the robust, robust data that is available in this, science community.

3:43:15
real quick, the water data library,

3:43:18
obviously just load it so you can actually load all of the data.

3:43:22
The stations for the water data library. Chart

3:43:26
if you zoom in to a particular area and you can chart this.

3:43:31
So this is looking at almost a year's worth of data.

3:43:34
From here you can zoom in and get more specific

3:43:40
and then start to add additional information on top of this.

3:43:44
So the QA/QC data is really important for those who are doing, reports

3:43:50
or studies or whatever.

3:43:53
The CDEC I know is not always what people are looking for

3:43:57
when they're trying to do, to,

3:44:00
peer reviewed work.

3:44:04
This just shows you, an example

3:44:06
of looking at CDEC and the water data library together

3:44:11
so you can go in and turn on the CDEC and then see the QA/QC differences.

3:44:18
It's a good way to just do a quick snapshot of of,

3:44:22
two very different data sets without having to download it.

3:44:26
Go find the station code, you know, all the hard work.

3:44:29
And then if you're satisfied with what you found

3:44:31
or you need to change your your start and end dates and then do another data grab

3:44:36
and then go ahead and download the information using the tools on the right.

3:44:43
the next

3:44:44
project that we're hoping to be, finished with in the next month

3:44:49
or so will be the EDSM Data Dashboards.

3:44:52
This will just give you an idea of what we've been working toward with U.S.

3:44:56
Fish and Wildlife.

3:44:58
so we're working to make it easy

3:45:02
for users to go in and pick in a particular time period.

3:45:05
So, like if you're doing today the current sampling week, previous sampling

3:45:09
week, a water year, etc., or you can set custom

3:45:13
time, it loads on the map

3:45:16
and so you'll get all of the fishery information on this map.

3:45:20
And each one of the colors represents a particular fish.

3:45:24
And then once that's loaded, you get all of the attribute information,

3:45:29
which is basically the monitoring surveys, results.

3:45:34
And you can click on the attribute table and zoom into those areas.

3:45:38
You can filter by species.

3:45:40
So, if you're looking for only delta smelt or a particular run of Chinook

3:45:45
salmon, you'll have the ability to update your map.

3:45:48
and the table with that information

3:45:52
and then start to get these snapshots, which are just widget badges

3:45:57
that will have, just highlights of important

3:46:00
information, like the total weekly Delta smelt catch,

3:46:05
winter Chinook runs catch, your highlights for the week

3:46:09
for a particular species and then charting all the other.

3:46:13
We're using other visualizations as well.

3:46:17
So the

3:46:19
the UI has many different ways of that

3:46:22
it's being laid out right now

3:46:23
based on the particular user group that's going to use it.

3:46:27
this is just the visualization of the map with a slide out over the top

3:46:32
so you can get your snapshot widgets, your weekly,

3:46:36
reports, your station surveys, your smelt totals and the data table.

3:46:41
So it's all in one UI.

3:46:43
And then obviously getting all of the important data and references,

3:46:47
links to the EDI, downloading the data

3:46:51
and then about the programs itself, a

3:46:55
quick snapshot of some visualization or ways of visualizing data

3:47:00
from this particular program.

3:47:06
And this is an,

3:47:08
looking at a particular report.

3:47:11
So this is where you can actually choose from a whole selection of, of saved maps

3:47:17
that are just set on a particular time interval, load those maps

3:47:21
and get all of the corresponding reporting widgets beneath it.

3:47:25
Tables, supporting everything that you see on the map.

3:47:31
Similar and some informational widgets.

3:47:36
So what the survey week was and the fork length ranges.

3:47:42
And then finally in Explore Data

3:47:45
where we were talking about how explore is more of this

3:47:48
free style location where you can go and just explore in any of the data sets

3:47:54
and then start to load them on the map and they add to the timeline below.

3:47:58
This is just an idea that we're working on right now of either

3:48:02
doing a quick view of a particular, data set that, you know, you want to look at,

3:48:06
or going in and doing a custom data load and then adding that on top of there.

3:48:12
So then finally, we have

3:48:16
from many years been requested from pretty much

3:48:20
most of the people in the community is how can I build my own dashboard?

3:48:24
Do you have so many data widgets from all of your dashboards?

3:48:28
throughout BDL.

3:48:29
And so we did develop a workspace builder that we all hope to release

3:48:33
come, late summer, early fall, where you can go in and start to build,

3:48:38
what you want, whether it's an operations widget or saved

3:48:42
maps, graphs that you need that you want to look at every day

3:48:47
and just have updated.

3:48:48
And so we've created a build widget

3:48:51
where you can go in and start just adding your widgets to a dashboard.

3:48:57
Find your your the type of widget that you're looking for,

3:49:00
whether it's mapping or its references or its species.

3:49:05
And you start to pull it all together and, yeah,

3:49:11
that'll

3:49:11
be a lot of fun for people who want to look at something very specific,

3:49:15
and it ends up in your profile page and it's all there for you.

3:49:19
So, here's the QR codes for the app and that's it.

3:49:25
I'm here for questions.

3:49:27
And so is Dave from 34 North. at the Q&A.

3:49:30
So thank you.

3:49:38
Thank you Amye, really cool presentation.

3:49:40
Thank you.

3:49:41
People are excited about some of your tools okay.

3:49:44
So we have a couple of minutes.

3:49:46
I just wanted to share a couple of visualizations

3:49:51
that some of us pulled together.

3:49:52
before we move on to the Q&A.

3:49:54
So this will be pretty informal and quick.

3:49:59
the background is that somebody, had requested

3:50:04
in our initial poll where we were, just seeking

3:50:09
feedback on what people wanted to learn about for this workshop.

3:50:12
so we had requested, having a figure, a gallery of commonly made figures.

3:50:17
and Trinh had mentioned that he was recently thinking about this

3:50:21
as a good idea as well.

3:50:23
So we thought we, would

3:50:27
just as we had time, stick a couple plots in a script and,

3:50:32
share that location in case it's helpful to anyone.

3:50:37
these are just a couple of ideas.

3:50:40
I think obviously if there is any feedback for requests,

3:50:45
or modifications, or if you want to add some of your own code that you use a lot,

3:50:50
happy to have you collaborate or make suggestions.

3:50:54
and I'll just walk through quickly some of the ideas that we had.

3:50:59
And it is in R markdown for now.

3:51:05
but it sounds like it would be easy to convert into Quarto as well.

3:51:10
So the link

3:51:12
let me go back and just stick this in.

3:51:15
I think I've already put this link in multiple times for other presentations,

3:51:19
but you can go there and then

3:51:22
in our GitHub page,

3:51:24
you'll see the initial code for a figure gallery as well.

3:51:30
And I just want to note that a lot of the scripts for actually

3:51:34
pulling data in from online for the visualization.

3:51:37
So like pulling data in from CDEC and,

3:51:42
Day Flow and things like that,

3:51:44
that's actually in a different script from what I'm going to be sharing.

3:51:47
But you can find all that in the GitHub.

3:51:50
Amye, can I just request you to mute?

3:51:57
Getting some background noise.

3:51:59
Thanks.

3:52:00
Okay.

3:52:01
So quick walkthrough.

3:52:03
this is another example of what

3:52:06
a an R markdown or Quarto can look like.

3:52:09
so sharing code, sharing,

3:52:14
sharing just ideas for visualizations.

3:52:16
When you're starting an analysis, I think it's a really good way to use

3:52:20
these tools.

3:52:21
So you can, you can hide the code if you want it to just be a smooth,

3:52:26
document that's not cluttered.

3:52:29
or you can show all the code, so that others

3:52:33
can just kind of copy and paste for these different figures.

3:52:37
So we have it organized here by some hydrological parameters

3:52:41
that are used a lot, and some environmental time series plots,

3:52:45
fish data and spatial data.

3:52:50
So, for

3:52:51
hydrological parameters, Rosie has some.

3:52:54
I think Rosie and Dave and maybe Trinh also have developed some code for pulling

3:52:58
in water year type and water year index

3:53:01
directly from the CDEC site.

3:53:05
So there's some code and then another script I showed you,

3:53:08
and then you could plot things like what are your type across here

3:53:12
for Sacramento Valley, San Joaquin Valley.

3:53:17
these are just some bar plots.

3:53:20
there's also some code for pulling,

3:53:23
delta outflow data directly from CDEC.

3:53:26
So this is a good source of data.

3:53:29
more in real time

3:53:32
since, day flow, is calculated later,

3:53:35
like after the year is over, and that data usually comes out a little bit later.

3:53:41
but yeah, there's code for downloading both of those

3:53:44
and then some examples for making plots, by year.

3:53:48
you can do two other things with it, of course.

3:53:52
a couple examples for how you would show multiple plots

3:53:55
so you could take, data

3:53:58
and kind of put it into long form and make a facet plot.

3:54:03
So these are

3:54:04
like these a little bit smoother, but you've got to rearrange the data to,

3:54:08
to make it fit your purposes and sometimes tweak axes and things like that.

3:54:12
so here's an example of that.

3:54:15
Or you can use packages a little bigger,

3:54:19
like patchwork or other packages that bind plots together,

3:54:22
where you can make two separate plots, one for outflow, one for export,

3:54:28
and then just with some simple code, put these

3:54:31
these two next to each other.

3:54:37
Okay.

3:54:37
So, environmental data, I think a lot of people work with this a lot.

3:54:41
So we've got some examples of pulling in multiples

3:54:45
stations with the same parameter or multiple parameters for the same station.

3:54:50
So here we're just looking at Rio Vista Bridge

3:54:54
data, and or providing

3:54:58
an example of plotting over the course of a year.

3:55:02
What, different parameters are looking like.

3:55:10
Just some example code

3:55:11
for how to export that into your files.

3:55:16
I think we've had some mention of some interactive plots

3:55:19
today, like Plotly, that Sam mentioned.

3:55:22
So here's just an example.

3:55:24
Like it's pretty similar to using ggplot.

3:55:27
Just use another function data in plotly to stick that in.

3:55:31
So here you can zoom in on data.

3:55:37
you can kind of find outliers if you see a little point

3:55:40
that you want to filter out, get the information on those.

3:55:49
bringing water year type in.

3:55:52
you can combine that with environmental data and,

3:55:56
look at environmental data by over several years

3:55:59
water year types.

3:56:05
cool trend plots

3:56:06
for comparing current data to historical data.

3:56:10
this code is from Mallard,

3:56:14
and I think it's called and you can look at all the other years

3:56:19
and then this year looking at water temperature there looks like we're

3:56:24
somewhat average ish, maybe a little bit above

3:56:26
average.

3:56:32
heat map plots.

3:56:34
You saw a couple examples that Susannah had on SacPas.

3:56:37
So this, just is an example of how you would make

3:56:40
that kind of plot in R.

3:56:43
really good way to look at multiple axes.

3:56:48
So year,

3:56:50
date and then, the actual temperature.

3:56:57
And this allows you to see,

3:56:59
really clearly that the hottest times of the year

3:57:02
is June through October.

3:57:06
Another really cool plot

3:57:07
that Trinh made that I didn't know about is this density ridge plot.

3:57:12
this allows you to see a whole distribution over time.

3:57:16
So this is showing of how the distribution of water temperature varies

3:57:22
by month and year.

3:57:30
And then we've got a couple of,

3:57:33
examples using Delta Juvenile Fish Monitoring.

3:57:37
Oh this one versus Delta smelt index.

3:57:39
So, we make these kinds of plots a lot, which is just plotting abundance over time.

3:57:44
One metric, this is the delta smelt index from Fall Midwater Trawl Survey.

3:57:49
just pulled directly from online.

3:57:54
And then these are some examples

3:57:55
from the Delta Juvenile Fish Monitoring Program.

3:57:59
again, a tile plot,

3:58:03
this is something I commonly make before

3:58:05
starting a report or study or something.

3:58:08
Just looking at, the sample sizes over time, over months.

3:58:13
this helps you kind of decide what you're going to include in your study,

3:58:17
make sure sample sizes are similar.

3:58:25
we've got some CPUE

3:58:26
plots, across here and,

3:58:30
this one's just filtered to a couple native fishes.

3:58:37
And then also looking over region.

3:58:40
I'm scrolling through this pretty fast.

3:58:43
But I'll let you all peruse.

3:58:48
here we have kind of a summary over time

3:58:51
of CPUE by water year and again visualized by water.

3:58:56
year type.

3:59:03
here's an example.

3:59:05
I think a lot of the time

3:59:06
we're doing some actions or like a restoration project or,

3:59:11
comparing some background

3:59:14
activity to, another variable.

3:59:17
So, this one is showing, inundation periods in the Yolo Bypass.

3:59:22
And then comparing that with chlorophyl a concentrations.

3:59:26
So these are useful just for showing

3:59:30
what is happening during an action or, or before an action or things like that.

3:59:40
And then, just

3:59:41
two quick examples for spatial data.

3:59:45
and a lot of people maybe learned arc map

3:59:47
or are working in Arc Pro or ArcGIS Pro.

3:59:53
but it is nice to be able to just code maps easily.

3:59:57
So some of us have converted over to R.

4:00:00
So here's basic code just for, converting your latitude

4:00:04
longitude data into an SF object, that allows you

4:00:09
it allows it to become like a spatial type of object.

4:00:13
And a lot of the time, the first thing I do is just look at the data.

4:00:17
this is an interactive display.

4:00:19
really quick, one line of code.

4:00:22
You can make this after you've converted your data and,

4:00:25
allows you again to hover and look at details of the data.

4:00:33
here we have also,

4:00:36
combined station with region.

4:00:39
these are EDSM regions, but DJFMP stations,

4:00:43
and you can just see kind of

4:00:46
these gray ones are the stations DJFMP samples, but not EDSM

4:00:51
and sorry for all the acronyms, I think we've heard them, but Delta

4:00:54
Juvenile Fish Monitoring Program and Enhanced Delta Smelt

4:00:58
Monitoring. So.

4:01:03
And then a static map.

4:01:04
I make this a lot for,

4:01:08
different studies when you're making a study map.

4:01:12
Study area map.

4:01:14
this is using,

4:01:16
base maps from Delta Mapper, which we learned about during Dave's talk.

4:01:20
this provides us with some nice regions

4:01:24
and a base map of the Delta.

4:01:27
And, we've got our, our sampling stations here in the middle.

4:01:33
Go North arrow and, scale bar.

4:01:38
So just basic map.

4:01:41
And that's all I have here.

4:01:43
Just wanted to walk through that quickly and show some of these examples.

4:01:46
And like I said, if there are others that want to add more

4:01:50
plots or, ask for more plots,

4:01:54
to be made, just reach out.

4:02:10
All right.

4:02:11
So that is the end of

4:02:15
our slotted talks today.

4:02:18
And we're going to

4:02:23
now move into Q&A.

4:02:25
I'm going to,

4:02:27
ask maybe all of the presenters come on camera.

4:02:31
I had a moment to, to do my talks with camera.

4:02:34
And I think we all

4:02:37
just, followed my example of not putting my camera on, but just so you can,

4:02:41
you can see our faces and, nice to see people respond to questions,

4:02:48
with their cameras on.

4:02:49
so I think this Q&A session is meant for anyone

4:02:54
to be able to answer questions, not just the presenters.

4:02:58
But I figured we have some people to look at right now.

4:03:01
And, depending on the question, we can,

4:03:05
see if other people in the audience have,

4:03:07
any comments or responses.

4:03:16
And I haven't seen any questions

4:03:19
that haven't been answered yet.

4:03:22
I'm going to open it up.

4:03:24
So first to.

4:03:28
To any of that before

4:03:31
hopping over to some of the ones we got from,

4:03:35
registration.

4:03:51
Also, if you have anything that you just want

4:03:53
to see in more detail or something that was touched on, but you want to,

4:03:58
like, look at the code again or something, feel free to ask for that as well.

4:04:08
Okay.

4:04:09
I'm going to just jump into something

4:04:12
and let people think and put things in the chat.

4:04:16
I think that most of the questions that got asked, it

4:04:20
through our registration form did get answered, but we can just

4:04:24
maybe make sure that everything's touched on.

4:04:26
And if whoever asked that question is here,

4:04:29
they can elaborate on their questions, if needed.

4:04:33
So the first one we got was, I'll pick up

4:04:38
post in the chat.

4:04:46
Please address integrating data from various agencies

4:04:50
and any plans to standardize and unify data

4:04:54
from various sources to streamline analysis.

4:05:03
Go ahead.

4:05:03
Rosie,

4:05:05
I can take a stab at this.

4:05:08
this is something that me, Dave. Sam.

4:05:12
Trinh a lot of people have put a lot of work into, and, it's.

4:05:21
I don't know,

4:05:21
in terms of, like, addressing, you kind of have to take what you get.

4:05:26
unfortunately, there's no magical way to get everyone to use the same column

4:05:32
names or the same format or, anything like that.

4:05:36
the data utilization workgroup does have, you know, some recommended,

4:05:41
vocabulary to use in your data sheets.

4:05:44
but a lot of it is dealing with each agency

4:05:50
has their own requirements, their own way of doing things, their own API's.

4:05:54
but a lot of those integrated data packages

4:05:57
that Dave mentioned during his talk have done the hard work for you.

4:06:03
They've got the code behind it, and they're pulling any new data

4:06:07
that gets out there in a regular basis.

4:06:10
so you can integrate them.

4:06:12
the data utilization workgroup has talked

4:06:15
many times about like, can we make any standards?

4:06:19
And the answer is we really can't.

4:06:20
if we tried, like we'd have

4:06:26
varying levels of,

4:06:29
people adopting them.

4:06:30
And if we make some sort of new standard that everyone changes everything.

4:06:33
And all the integrated data sets we've already made will break. So,

4:06:39
I think there's a lot

4:06:40
the biggest thing that we have succeeded in doing, to a large extent,

4:06:44
is getting everyone to publish their data in open, easy to access format.

4:06:51
And exactly what that format is matters less than that.

4:06:54
It's out there publicly available, which is, you know, the best.

4:06:58
So, I don't know if Dave or anyone else, Trinh

4:07:01
wants to add to that.

4:07:06
Yeah.

4:07:07
So I agree, it's really hard to have everyone

4:07:10
adopt the same standard, even though that's ideal.

4:07:13
but Rosie did mention that we've been able to integrate

4:07:18
all these data sets together already in these packages.

4:07:21
And so that says that we can do it.

4:07:24
And one way we been able to do it is to use a crosswalk.

4:07:28
and so

4:07:29
we could theoretically work on a crosswalk while allowing everyone to remain

4:07:33
using the, their, their current column names and etc.

4:07:38
So that's one way we can do it.

4:07:39
But to have that entire system adopt a standard

4:07:43
might be a really hard task to do.

4:07:46
I wish we could though. So

4:07:49
I can dream.

4:07:59
Denise.

4:07:59
Go ahead.

4:08:01
one thing I mentioned in my presentation, too,

4:08:04
I think that could potentially help is integrating

4:08:07
data, integrating preliminary data at an early stage.

4:08:11
And maybe that data isn't something that you could use for a peer review

4:08:14
publication or something like that,

4:08:16
but it is something that will help people get a bigger picture of everything.

4:08:19
That's going on earlier.

4:08:21
and also, it could help, different programs

4:08:26
figure out what isn't fitting together with their data.

4:08:30
we've we've done a little bit of that through the Delta Smelt supplementation team.

4:08:34
And found a lot of different things that we could, work on

4:08:38
to just make that a little bit easier to integrate in the future.

4:08:41
So, like I said, it may not be that useful for a final peer

4:08:45
reviewed product, but it does help get everyone on the same page

4:08:49
get a better, bigger picture of what's going on earlier in the process To.

4:08:58
Do you have any ideas for

4:08:59
how to make that happen, or have you tried?

4:09:04
right now we're just using Excel spreadsheets

4:09:07
on SharePoint, and there's probably much more sophisticated ways to do that.

4:09:12
it's kind of serving the purpose for now, but,

4:09:16
I'd be I'd be happy to hear any recommendations for making that better.

4:09:21
But the the shared spreadsheet serves a purpose for now.

4:09:24
It's just a lot of coordination and communication.

4:09:34
And just wanted to show really quickly.

4:09:36
I pulled up the crosswalk.

4:09:38
so people can know what we're talking about.

4:09:40
we had this effort

4:09:45
within the data publication workgroup of,

4:09:49
just providing some suggestions for,

4:09:53
variable names.

4:09:55
this was more directed at people who are maybe starting a new study,

4:09:59
or were in the process of updating databases or things like that.

4:10:05
because, yeah, it can be very hard

4:10:07
to change all of your variables, if you already have something going.

4:10:10
But yeah, we did make some suggestions for how to name things

4:10:13
and what kinds of variables to include and units and things like that.

4:10:17
as well as a, like a fish code

4:10:20
crosswalk that at least makes it easier to, to bring fish together.

4:10:24
So we're hoping people could include that code as well as use their own code

4:10:28
if they had another one.

4:10:29
but that when people doing synthesis or bring data

4:10:32
together, it would just be one step easier.

4:10:36
So I know a couple of programs started, including that in their published

4:10:39
data sets.

4:10:46
Okay.

4:10:46
Do we want to let's jump into some of

4:10:51
the questions that have come up.

4:10:52
There's a question for Trinh.

4:10:56
I'll read it and then let Trinh respond.

4:10:59
You mentioned including metadata for historical data, but what wasn't

4:11:02
discussed was updating the older data to bring it up to current data standards.

4:11:07
Is anyone doing this?

4:11:11
Yeah.

4:11:12
So that's a really good question.

4:11:14
we started looking at historical metadata not too long ago, actually.

4:11:21
because like I've said in my presentation,

4:11:23
we've always really had a bias towards doing the current metadata.

4:11:27
and so when you start to go down the rabbit

4:11:31
hole of things like machine learning and things like that, where you can

4:11:34
you start using big data, you, you start wanting to use more data.

4:11:38
And so that's why we started talking about all this stuff,

4:11:41
and trying to make our entire data set usable.

4:11:44
And so we, the Data Publication working Group have put together

4:11:49
a template, a sort, of things that we think

4:11:54
you should try to document,

4:11:58
for your historical metadata to.

4:12:01
And that is similar to a metadata standard.

4:12:04
we didn't really follow a metadata standard

4:12:07
organization as much, but it's in a nested list format.

4:12:12
And so that guide is out there.

4:12:14
and that's one way you can perhaps start

4:12:18
recording the historical metadata related to your, your data set.

4:12:22
but for your question about is anyone doing this?

4:12:27
we are trying,

4:12:31
trying to work through a lot of the older data

4:12:35
and seeing if there were changes to the sampling protocols that has happened

4:12:41
since using that historical metadata template that I mentioned.

4:12:46
But it's pretty difficult, when you've had decades worth of data

4:12:51
and you've had many different survey leads

4:12:55
that have come and gone, it's really difficult

4:12:58
to try and go back in time and figure out what they did.

4:13:04
and so we

4:13:06
can try the best we can, but there's going to be a point where it's

4:13:10
just not possible to really describe what happened in the past.

4:13:15
That's not to say, though,

4:13:17
that we shouldn't be doing that now.

4:13:20
We shouldn't be to doing that now because what is current metadata

4:13:24
now is going to be historical metadata later.

4:13:26
and so we've been pushing people to try

4:13:29
and document the historical metadata to the best of their abilities.

4:13:34
And if they're not able to, well,

4:13:37
at least have that in mind now and, and continue to

4:13:41
to document as much of the metadata as possible for the future.

4:13:48
Next. Cat.

4:13:48
Yeah.

4:13:48
So this is the the template that we put together.

4:13:53
It has all the general elements, that we think you should

4:13:57
consider, describing for your data set.

4:14:01
and then we also have various examples in there

4:14:04
that hopefully will make it a bit easier for you to follow.

4:14:08
I think there's three types here that we described.

4:14:11
there was fish data, water quality data and the zooplankton data

4:14:16
or something like that.

4:14:17
But hopefully

4:14:21
it's logical enough where,

4:14:23
it should be fairly easy to follow.

4:14:33
Yeah.

4:14:34
And I know that when we like at DWR, when I was there, there,

4:14:38
we had created some historical meta data

4:14:41
templates as well for like, continuous data sensors and things like that.

4:14:44
And, programs were starting to implement that.

4:14:48
so there are people, that

4:14:51
know that it's important and are doing it.

4:14:54
And I think it helps to have, examples and, and templates to follow

4:14:59
to know what needs to be there.

4:15:03
and like, I know one example to like Sam brought together a lot of these

4:15:07
synthesis data sets, and had and really needed information on

4:15:13
how sam how lab procedures

4:15:16
changed with like, zooplankton sampling in order to,

4:15:19
make an effective data set and,

4:15:22
that was where one of our programs realized, like,

4:15:26
really, we really needed to document that and keep track of it

4:15:29
in going down the line so that people could include

4:15:32
the data set in products like that.

4:15:35
So just an example of the importance of it.

4:15:38
don't always realize it until somebody tries to use your data later on.

4:15:53
Okay.

4:15:53
There's.

4:15:56
Another question.

4:16:00
A curmudgeonly question.

4:16:02
Can you speak about mitigating the potential downsides of open data?

4:16:07
You did a great job discussing the many broad societal benefits,

4:16:10
but I'm thinking in particular about students and early career scientists.

4:16:14
This is maybe less relevant for public agency scientists.

4:16:19
How do these folks avoid getting scooped on their hard earned data

4:16:23
without waiting until after publications

4:16:26
to publish data?

4:16:35
Rosie.

4:16:38
I hate to always be the first one to jump in,

4:16:40
but I have strong opinions about this.

4:16:42
when was the last time someone actually

4:16:46
picked up a grad students that wasn't their professor,

4:16:50
picked up a grad students work and published on it,

4:16:54
before they could.

4:16:57
I feel like it's a it's

4:16:58
something that there's a lot of fear over.

4:17:01
But there are so many ways to analyze that data

4:17:06
set. The chance of two independent researchers

4:17:09
wanting to do the exact same thing with the exact same data is pretty small.

4:17:13
That being said, there's absolutely nothing wrong with having,

4:17:16
you know, a year or two between collecting the data and publishing it,

4:17:20
especially if you're a student or early career researcher.

4:17:23
before, you know, making it totally open.

4:17:26
I think even if you are

4:17:29
an early career researcher, if someone contacts you and says, hey,

4:17:32
you know, can I use the data you collected for a reason a to do thing

4:17:36
B that's completely not what you were planning.

4:17:40
Yeah.

4:17:40
Like, let them have it.

4:17:42
The more the merrier.

4:17:45
My thoughts anyway.

4:17:51
Yeah,

4:17:51
I think that is one of the concerns that I've heard most about.

4:17:55
And yeah, as people who work for agencies like it's it's public data anyway.

4:18:00
So the no problem with posting it as soon as possible and it's, it's required.

4:18:07
Right.

4:18:07
And any studies even university studies that are funded by government agencies

4:18:12
I think have requirements as well to post within a certain amount of time.

4:18:17
but for people who aren't.

4:18:19
Yeah.

4:18:19
Like Rosie said, I think it's pretty common

4:18:21
to have a one year or two year holding period for your data,

4:18:27
before publishing.

4:18:27
And I think that's the practice

4:18:29
for some of the products that I've been hearing about.

4:18:34
but yeah, I think getting into the practice

4:18:38
of putting data up, like some of those benefits,

4:18:41
getting it up as a student,

4:18:43
you know, can give you more citations and things like that.

4:18:47
and make connections since you can share your data with others and collaborate.

4:18:51
So there are benefits there too.

4:18:54
But I don't know if there's anyone else in the audience

4:18:56
who has,

4:18:59
has worked with this issue

4:19:00
or has any thoughts on that.

4:19:07
I just wanted to say that

4:19:08
this has actually been a topic discussed by the FAIR principles.

4:19:12
so it's a common enough concern that

4:19:16
the accessibility portion of FAIR does touch on this.

4:19:21
And, does state, need

4:19:24
at the very least make your metadata available.

4:19:27
and so if you can at least make your metadata available,

4:19:31
people know who to, contact, who to cite for certain things.

4:19:37
And if they want new data, they can then reach out

4:19:39
and know all the channels required to access that data.

4:19:42
And so, having authorization to,

4:19:47
to access a database is also part of the FAIR principle.

4:19:51
and for some sort of data sets that it is appropriate.

4:19:55
And so if you feel that that is appropriate for your data set,

4:19:58
at least make your metadata available at that point.

4:20:07
Thanks Trinh.

4:20:12
All right.

4:20:12
Next question I've got is all these websites shiny apps dashboards are great.

4:20:18
Do you folks with CDFW or other state

4:20:20
agencies run into issues within their department

4:20:24
with getting approval for creating these external websites?

4:20:28
Could you speak to the process?

4:20:31
I'm new and I've been told changing anything on our website

4:20:33
is difficult.

4:20:42
let's go ahead Sam.

4:20:45
Yeah, I think Rosie

4:20:47
partially responded in the chat.

4:20:50
I mean, every department handles the website a different way,

4:20:54
so it kind of depends on your departmental policy, I guess.

4:20:57
I wouldn't know how CDFW handles it.

4:21:00
And but that really usually just applies to the like for state agencies,

4:21:04
the .ca.gov, the official government websites.

4:21:08
And usually if you're publishing something on another domain,

4:21:11
if it's a GitHub page or a shiny aps IO site,

4:21:17
you know, you have a little bit more discretion to publish things

4:21:20
without it going through multiple rounds of review or through the chain.

4:21:24
I mean, of course, you should always clear things

4:21:27
with your supervisor before you make it publicly available,

4:21:31
but that's kind of one of

4:21:32
the advantages of these external websites is that you have a little bit

4:21:37
more freedom to do it yourself, rather than having to work within

4:21:40
your internal bureaucracy of how things

4:21:43
get on websites.

4:21:52
Yeah, I was like part of the oh sorry Trinh, go ahead.

4:21:54
That's

4:21:56
I was going to have.

4:21:57
That's part of the reason

4:21:58
I like created a story map was it was a easier user interface

4:22:02
for me to update and share with folks and keep it up to date than going through

4:22:07
our website on like .ca.gov, but had a lot more

4:22:12
like hurdles to jump over to make any updates.

4:22:16
So, it provided a little more flexibility

4:22:19
there, which is convenient.

4:22:23
Yes, I agree with all of that.

4:22:25
I think using the third party websites are useful, but,

4:22:29
I just wanted to touch a little bit on the CDFW portion of that

4:22:32
because I've ran into these issues already.

4:22:35
when you get going with coding and things like that, especially with scraping,

4:22:41
using R to populate perhaps data for some of these apps.

4:22:47
CDFW IT does have some firewall issues.

4:22:50
And so, if you do have that issue, you do have to reach out to ITy.

4:22:55
There are solutions that they have,

4:22:57
but you do have to first reach out to them and they'll guide you through it.

4:23:12
All right.

4:23:13
Thanks everyone.

4:23:13
I don't see

4:23:18
any other new questions.

4:23:20
So we have a few minutes left.

4:23:23
I'll put in another

4:23:28
question from the

4:23:31
registration.

4:23:36
This one is,

4:23:37
about data storage and transfer.

4:23:40
is there a link formula for data storage and transfer,

4:23:44
or do we need to accept multiple specialized formats?

4:23:51
I think I know Rosie touched on some of this,

4:23:54
and I think you touched on some of this in your talk.

4:23:58
as to, like, formatting your data and what is recommended.

4:24:04
I think the general

4:24:09
preferred, guidelines, like we were talking

4:24:14
about accessibility and things like that is to have open source formats.

4:24:17
So like csv's that anyone can access

4:24:21
without having specialized software.

4:24:27
But there are some, some software that are required

4:24:31
for certain types of data or methods that people want to use.

4:24:36
that just require a different format.

4:24:37
But, hopefully when sharing data, if it's possible to convert

4:24:42
whatever special file it is, into something that others can access and use,

4:24:46
like for example, converting the data to a CSV or something else.

4:24:51
that would be again, preferable.

4:24:53
I think it it depends on the situation.

4:24:57
And if there's more specificity in that question that I didn't cover,

4:25:01
let me know if there's another question there

4:25:04
or if anyone else wants to respond to that.

4:25:06
Go ahead.

4:25:23
There's

4:25:23
another similar question about, do we have preferred platforms for data

4:25:27
hosting, publishing, sharing and storage?

4:25:31
So, in terms of that

4:25:35
just covers a couple of, of extra things of

4:25:38
so for hosting and publishing I think we use EDI and FTP's

4:25:42
a lot from what I've seen.

4:25:45
and like you've heard, Trinh has talked about all the benefits of EDI.

4:25:51
so I think for, you know, cleaned up data

4:25:54
and on kind of an annual basis, we like to host our data there.

4:25:59
but I've heard from some of the, the SacPas people like Matt,

4:26:02
who manages the database that, some of the other forms

4:26:06
can be kind of easier to see updates from as well,

4:26:10
like the FTP's update more readily, things like that.

4:26:14
So, so some people use those and that's good for for data coming in.

4:26:18
on a more regular basis.

4:26:21
for sharing

4:26:24
data, I think you've seen like SacPas and Bay Delta Live are really great places

4:26:28
to share your data in, like a visual form,

4:26:32
or to to get your data quickly up somewhere that people can, can access,

4:26:36
like a central place for people to find it.

4:26:41
And for storage,

4:26:42
I don't think we have I don't know if anyone has recommendations

4:26:47
for how to store your data.

4:26:48
I think I had covered, like archiving is really important for storage. And

4:26:52
a lot of people use their network drives at their agencies, which have backups.

4:26:57
I don't know if anyone uses like, any cloud kind of I mean,

4:27:00
like Microsoft has its own cloud.

4:27:05
what is it called?

4:27:06
OneDrive, that kind of thing.

4:27:08
But, yeah, I'm curious if anyone else has recommendations

4:27:12
for for any of those or storage in particular, because I'm not sure.

4:27:19
I wanted to jump in again.

4:27:21
just making sure that, like, I think a lot of people don't

4:27:26
necessarily understand the difference between storage

4:27:28
and archiving because they sound kind of the same.

4:27:31
Archiving is like EDI,

4:27:34
the perfect, beautiful data set in long term storage.

4:27:38
Storage is just, you're working

4:27:41
copy that you're updating all the time.

4:27:43
And so, yeah, most people have it on, you know, a local server

4:27:47
through their agency or, and, or, some kind of cloud network.

4:27:53
DWP has switched over

4:27:56
to OneDrive for a lot of, especially smaller databases.

4:28:00
so yeah, something like that.

4:28:03
But you want to have it backed up in multiple places potentially.

4:28:08
So, it'll be the DWRs OneDrive is

4:28:12
also synched to DWR servers periodically.

4:28:17
So on the cloud, and I'll have a local copy on my desktop.

4:28:21
So, yeah.

4:28:22
Then archiving in those long-term data repositories like EDI.

4:28:29
And as Rachel mentioned,

4:28:31
to be compliant with AB 1755,

4:28:35
the open and transparent water data, act, it needs to also be linked

4:28:40
on the CNRA platform, which I put the link in the chat.

4:28:53
Okay.

4:28:53
Well, I think we're getting close to the end.

4:28:56
it is 328 and we are ending at 330.

4:29:00
Thanks for sitting through this day.

4:29:02
We meant to make it four hours and keep it short,

4:29:06
but there was so much information and, interest in different topics.

4:29:09
So thanks for sticking through this.

4:29:11
Hopefully the breaks helped you, have a little time to

4:29:16
to rest and still absorb everything.

4:29:19
I'm going to turn it over to Dave for our final closing, to share out, for

4:29:24
the future steps.

4:29:31
Alright.

4:29:35
Just loading a couple slides here.

4:29:40
All right.

4:29:43
so. Yeah, first of all, thanks for coming.

4:29:47
I, thought that worked out really well.

4:29:49
It was a really great workshop.

4:29:51
I hope you guys learned a bunch of, of new information

4:29:54
and got a bunch of good links to, to go back into your own work

4:29:58
and start incorporating some of this into your work.

4:30:01
so first I just want to give, some thank yous, here.

4:30:06
So, Yeah, I want to thank the workshop organizers.

4:30:09
Cat, I mean, you did the bulk of the work to organize this workshop, so,

4:30:14
I really want to give a shout out to you, others that also did

4:30:18
a lot of work to organize this, for, Rosie and Sam and Trinh.

4:30:23
so thank you all for, for, for doing that.

4:30:26
Also, I want to thank all the great presenters

4:30:30
today, that there were some really great talks.

4:30:32
And so thank you for it.

4:30:33
Thank you for spending your time to prepare your talks and to come here

4:30:36
and provide your, your knowledge and information to the, to the group.

4:30:41
And then I also, of course, just want to thank all of you, all the attendees,

4:30:44
for coming in for your interest and your engagement here.

4:30:47
So, I mean, you guys are what keep us doing these kind of things.

4:30:50
So, if it wasn't for you, we wouldn't be here.

4:30:52
So thanks for for coming.

4:30:58
So yeah, just kind of.

4:31:00
We already kind of talked about this.

4:31:02
but, we're going to send out some resources after this workshop.

4:31:07
and the main, the main place that you're going to find a lot of our

4:31:10
resources are on this, workshop GitHub repo.

4:31:14
and we will I mean, I think we put that chat, that link in the chat, like,

4:31:18
many, many times, but we'll send it out in an email,

4:31:21
as a kind of recap email to everyone who attended.

4:31:25
and so in that, in that GitHub, you'll find, some example code,

4:31:30
you'll find, I think that figure gallery that Cat went over, like pretty much

4:31:35
all the presentations, that we had today and resource list

4:31:39
that has a whole bunch of really useful links for you all.

4:31:43
And we're also going to send out the, the workshop recording.

4:31:47
So if you missed anything or if you want to go back to something,

4:31:50
you'll have access to that as well.

4:31:52
And then another thing we're going to send out to you all is a post workshop

4:31:56
poll, to just kind of get, you know, some feedback

4:31:58
about how today's workshop went and then also to kind of look for some,

4:32:02
ideas for some future, ideas for training and such.

4:32:06
So, that's all I really have.

4:32:08
I don't know if any of the other organizers

4:32:09
have anything to add to that, but, if not, then we could,

4:32:13
give you the rest of your afternoon.

4:32:17
Yeah, just thank you, everyone.

4:32:19
And, Yeah, have a good rest of your day.

4:32:22
And, along with all those resources, I will get the recordings

4:32:27
uploaded as soon as possible and send out a link to those as well,

4:32:32
and sort of see.

4:32:34
Thank you. Well,

4:32:36
thanks everyone.