Propuesta de estructura de la base de datos para la plataforma de calidad del agua.

(enero de 2024, Fidel Serrano Candela y Abraham Toriz Cruz)

Esta propuesta de estructura de base de datos está basada en las muestras proporcionadas por la FGRA, en conjunción con el trabajo previo del equipo del LANCIS. El primer resultado es la lista exhaustiva de las variables que aparecen en las muestras proporcionadas en combinación con las variables de las fichas desarrolladas en colaboración entre LANCIS y la FGRA.

Adicionalmente se realizó un análisis de las ventajas y desventajas de diversas estructuras de datos y se determinó la mejor opción para este proyecto.

Análisis de bases de datos existentes.

Se analizaron los ejemplos de bases de datos existentes proporcionados que fueron 5 archivos en formato Excel y 4 archivos en formato pdf. Se utilizaron técnicas de análisis de datos para automatizar parcialmente la extracción de las variables de cada ejemplo, para la conformación de la lista exhaustiva de variables encontradas.

Lista exhaustiva de variables

A continuación, se listan las variables recolectadas. Las columnas de nombres distintos con el mismo tipo de información se reducen y normalizan.

Datos

#	Variable	Unidad	Columna normalizada
1	Amonio	mg/L	amonio
2	Arsénico	mg/L	arsenico
3	Cadmio	mg/L	cadmio
4	Cafeína	na	cafeina
5	Cloro	mg/L	cloro
6	Clorofila a	mg/m3	clorofila_a
7	Cobre	mg/L	cobre
8	Colformes fecales	UFC/100 mL	colformes_fecales
9	Color verdadero	“nm, UPtCo”	color_verdadero

#	Variable	Unidad	Columna normalizada
10	Conductividad eléctrica	mS/cm = 1mmho/cm	conductividad_electrica
11	Cromo	mg/L	cromo
12	Demanda Bioquímica de Oxígeno	mg/L	demanda_bioquimica_de_oxigeno
13	Demanda Química de Oxígeno	mg/L	demanda_quimica_de_oxigeno
14	Derivados de combustibles	mg/L	derivados_de_combustibles
15	Derivados de solventes	mg/L	derivados_de_solventes
16	Dureza	mg/L	dureza
17	Enterococos fecales	UFC/100 mL	enterococos_fecales
18	Escherichia coli	UFC/100 mL	escherichia_coli
19	Fierro	mg/L	fierro
20	Fluoruro	mg/L	fluoruro
21	Fósforo	mg/L	fosforo
22	Giardia lamblia	quistes/20L	giardia_lamblia
23	Grasas y aceites	mg/L	grasas_y_aceites
24	Material flotante	NA	material_flotante
25	Mercurio	mg/L	mercurio
26	Microcystina-LR	mg/L	microcystina_lr
27	Niquel	mg/L	niquel
28	Nitratos	mg/L	nitratos
29	Nitritos	mg/L	nitritos
30	Nitrogeno amoniacal	mg/L	nitrogeno_amoniacal
31	Nitrógeno	mg/L	nitrogeno
32	Ortofosfatos	mg/L	ortofosfatos
33	Oxígeno	mg/L	oxigeno
34	PH	1-14	ph
35	Plomo	mg/L	plomo
36	Saam	mg/L	saam
37	Salinidad	% ó ppt	salinidad
38	Silice reactiva	mg/l	silice_reactiva
39	Sulfatos	mg/l	sulfatos
40	Sólidos disueltos	mg/L	solidos_disueltos
41	Sólidos suspendidos	mg/L	solidos_suspendidos
42	Temperatura	°C	temperatura

#	Variable	Unidad	Columna normalizada
43	Trihalometanos	mg/L	trihalometanos
44	Trix	na	trix
45	Turbidez	UFN	turbidez
46	Zinc	mg/L	zinc

Metadatos

Información encontrada que describe la muestra, sin ser información cuantitativa sobre la misma.

Variable	Tipo de dato	Descripción
ubicacion	geometry(point)	Referencia geográfica del lugar de la muestra
fecha	timestamp	fecha en que fue tomada la muestra (con hora)
sitio	string	nombre del sitio de muestreo
ecosistema	string	ecosistema del sitio donde se colectó la muestra
color	string	ver documento a.0442_estcalidadagua.pdf
olor	string	ver documento a.0442_estcalidadagua.pdf
sabor	string	ver documento a.0442_estcalidadagua.pdf
fuente de abastecimiento	string	
usos del agua	string	
condiciones de la fuente de abastecimiento	string	
responsable	string	persona que tomó la muestra

Datos no reconocidos

Se encontraron en los ejemplos, pero se desconoce su significado.

- orp (potencial óxido reducción)
- cla (µg/l) (puede ser clorofila A, pero no coincide la unidad)

Estructuras potenciales para la base de datos

Formato wide

La información se almacena en una sola tabla (llamémosla sample) cuyos registros representan una muestra cada uno, con columnas describiendo cada valor tomado de la muestra (por ejemplo temperatura, concentración de oxígeno, PH, etc.)

id	timestamp	metadata1	ph	nitratos	...
1	2023-01-24 11:53:12Z	val1	6	0.0012	...
2	2023-01-24 11:55:12Z	val2	7	0.0071	...
3	2023-01-25 11:55:12Z	val3	5	0.0004	...

Ventajas

- Cada renglón es una unidad geográfica con toda la información de los parámetros que se midieron, y en ese sentido es mas directa su representación geográfica.
- Las consultas para cálculos sobre las muestras son directas.

Desventajas

- Si las muestras solo miden algunos parámetros, la tabla estará llena de valores nulos de los parámetros que no se midieron

Formato long

Se tiene una tabla sample cuyos registros tienen solo características de identificación del mismo y algunos metadatos (e.g. fecha de la muestra) y los valores de la muestra se guardan en una segunda tabla values con las columnas sample_id, variable, value y cada renglón representa un valor de la muestra (e.g. concentración de oxígeno o PH)

Tabla sample

id	timestamp	metadata1
1	2023-01-24 11:53:12Z	val1
2	2023-01-24 11:55:12Z	val2
3	2023-01-25 11:55:12Z	val3

Tabla value

sample_id	variable	value
1	pH	7

sample_id	variable	value
1	temperatura	22.2
2	pH	5
2	nitritos	33.3

Ventajas

- En el caso de información dispersa solo se guardan valores para variables de las que se tiene registro, ahorrando espacio de almacenamiento.

Desventajas

- Las consultas para cálculos sobre los valores de las muestras son indirectas.
- Se debe tomar una decisión general sobre el tipo de dato con el que se almacenan los valores puesto que todos los registros deben tener el mismo tipo de dato en la tabla values. Pero parece natural que el tipo de dato sea de punto flotante pues salvo el ecosistema, el color, olor y sabor todas las variables que se miden se reportan en punto flotante.

Columna JSON

Se tiene una sola tabla sample cuyos registros representan una muestra y los valores de la misma se almacenan en una única columna (todos) de tipo JSON serializados.

id	timestamp	metadata1	values
1	2023-01-24 11:53:12Z	val1	{"ph": 7, "nitratos": 0.15}
2	2023-01-24 11:55:12Z	val2	{"nitratos": 15, "concentracion_oxigeno": .023}

Ventajas

- Cada valor de la muestra puede tener su propio tipo de dato
- No se desperdicia espacio de almacenamiento pues solo se almacenan valores capturados

Desventajas

- Existe cierta indirección para utilizar los valores para cálculos sobre las muestras pues es necesario deserializar el JSON (sin embargo es importante mencionar que la base de datos ofrece facilidades para hacer esto pues tiene soporte completo de columnas JSON)
- Si la información no es dispersa este tipo de almacenamiento resulta de hecho más pesado que la alternativa de tabla expandida variable, unidades

Estructura seleccionada para la base de datos de la plataforma

La estructura seleccionada fue formato wide, con toda la información de una toma de muestra en un mismo renglón.

id	timestamp	metadata1	ph	nitratos	…
1	2023-01-24 11:53:12Z	val1	6	0.0012	…
2	2023-01-24 11:55:12Z	val2	7	0.0071	…
3	2023-01-25 11:55:12Z	val3	5	0.0004	…

La razón para seleccionar esta opción es que el manejador de bases de datos Postgres-Postgis, desarrolló una estrategia para no usar memoria en valores nulos lo cual redunda en que la desventaja de esta opción queda eliminada y se conservan las ventajas. Los renglones en la base de datos contendrán tanto los datos como los metadatos. A continuación se enlistan las columnas que conformarán cada renglón.

Metadatos

Variable	Tipo de dato	Descripción
ubicacion	geometry(point)	Referencia geográfica del lugar de la muestra
fecha	timestamp	fecha en que fue tomada la muestra (con hora)
sitio	string	nombre del sitio de muestreo
ecosistema	string	ecosistema del sitio donde se colectó la muestra
color	string	ver documento a.0442_estcalidadagua.pdf
olor	string	ver documento a.0442_estcalidadagua.pdf
sabor	string	ver documento a.0442_estcalidadagua.pdf
fuente de abastecimiento	string	
usos del agua	string	
condiciones de la fuente de abastecimiento	string	
responsable	string	persona que tomó la muestra

Datos

#	Variable	Unidad	Columna normalizada
1	Amonio	mg/L	amonio
2	Arsénico	mg/L	arsenico
3	Cadmio	mg/L	cadmio
4	Cafeína	na	cafeina
5	Cloro	mg/L	cloro
6	Clorofila a	mg/m3	clorofila_a
7	Cobre	mg/L	cobre
8	Colformes fecales	UFC/100 mL	colformes_fecales
9	Color verdadero	“nm, UPtCo”	color_verdadero
10	Conductividad eléctrica	mS/cm = 1mmho/cm	conductividad_electrica
11	Cromo	mg/L	cromo
12	Demanda Bioquímica de Oxígeno	mg/L	demanda_bioquimica_de_oxigeno
13	Demanda Química de Oxígeno	mg/L	demanda_quimica_de_oxigeno
14	Derivados de combustibles	mg/L	derivados_de_combustibles
15	Derivados de solventes	mg/L	derivados_de_solventes
16	Dureza	mg/L	dureza
17	Enterococos fecales	UFC/100 mL	enterococos_fecales
18	Escherichia coli	UFC/100 mL	escherichia_coli
19	Fierro	mg/L	fierro
20	Fluoruro	mg/L	fluoruro
21	Fósforo	mg/L	fosforo
22	Giardia lamblia	quistes/20L	giardia_lamblia
23	Grasas y aceites	mg/L	grasas_y_aceites
24	Material flotante	NA	material_flotante
25	Mercurio	mg/L	mercurio
26	Microcystina-LR	mg/L	microcystina_lr
27	Niquel	mg/L	niquel
28	Nitratos	mg/L	nitratos
29	Nitritos	mg/L	nitritos
30	Nitrogeno amoniacal	mg/L	nitrogeno_amoniacal
31	Nitrógeno	mg/L	nitrogeno
32	Ortofosfatos	mg/L	ortofosfatos

#	Variable	Unidad	Columna normalizada
33	Oxígeno	mg/L	oxigeno
34	PH	1-14	ph
35	Plomo	mg/L	plomo
36	Saam	mg/L	saam
37	Salinidad	% ó ppt	salinidad
38	Silice reactiva	mg/l	silice_reactiva
39	Sulfatos	mg/l	sulfatos
40	Sólidos disueltos	mg/L	solidos_disueltos
41	Sólidos suspendidos	mg/L	solidos_suspendidos
42	Temperatura	°C	temperatura
43	Trihalometanos	mg/L	trihalometanos
44	Trix	na	trix
45	Turbidez	UFN	turbidez
46	Zinc	mg/L	zinc

En el proceso de carga de las bases existentes es posible que tanto la lista de metadatos como la lista de variables crezcan, pero esta es la propuesta inicial para la estructura de datos de la plataforma de calidad del agua.