TableRepoViz

Visualizing Tabular Data Repositories for Facilitating Descriptive Tag Augmentation

CPSC 547 Project
Jianhao Cao

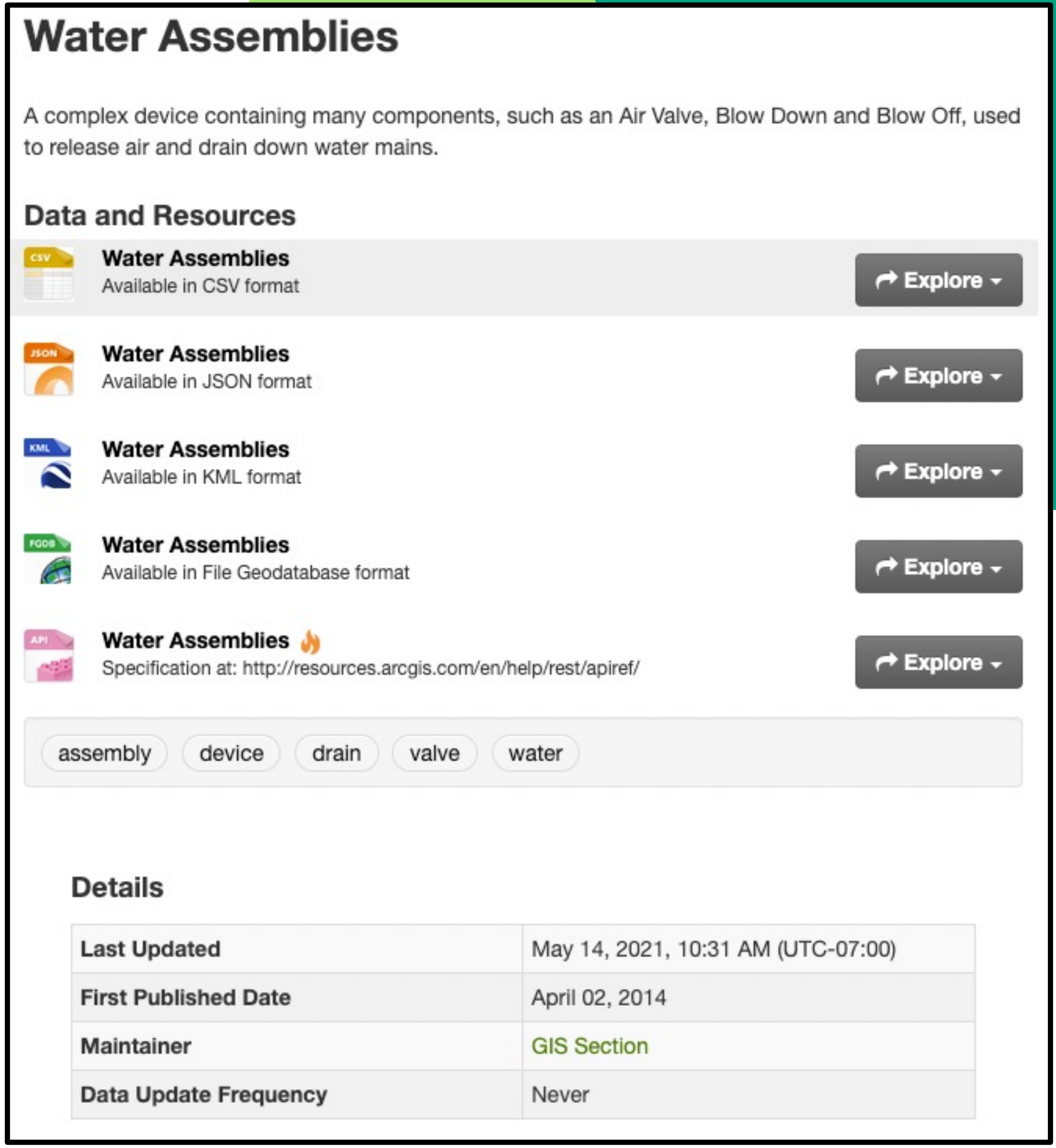
Background

Why tags in a table repo?

- Descriptive metadata for the dataset
- Functional demand for data search

Tag augmentation requirements

- efficiency
- accuracy
- consistency



Water Assemblies

A complex device containing many components, such as an Air Valve, Blow Down and Blow Off, used to release air and drain down water mains.

Data and Resources

Water Assemblies
Available in CSV format

Water Assemblies
Available in JSON format

Water Assemblies
Available in KML format

Water Assemblies
Available in File Geodatabase format

Water Assemblies
Specification at: http://resources.arcgis.com/en/help/rest/apiref/

assembly device drain valve water

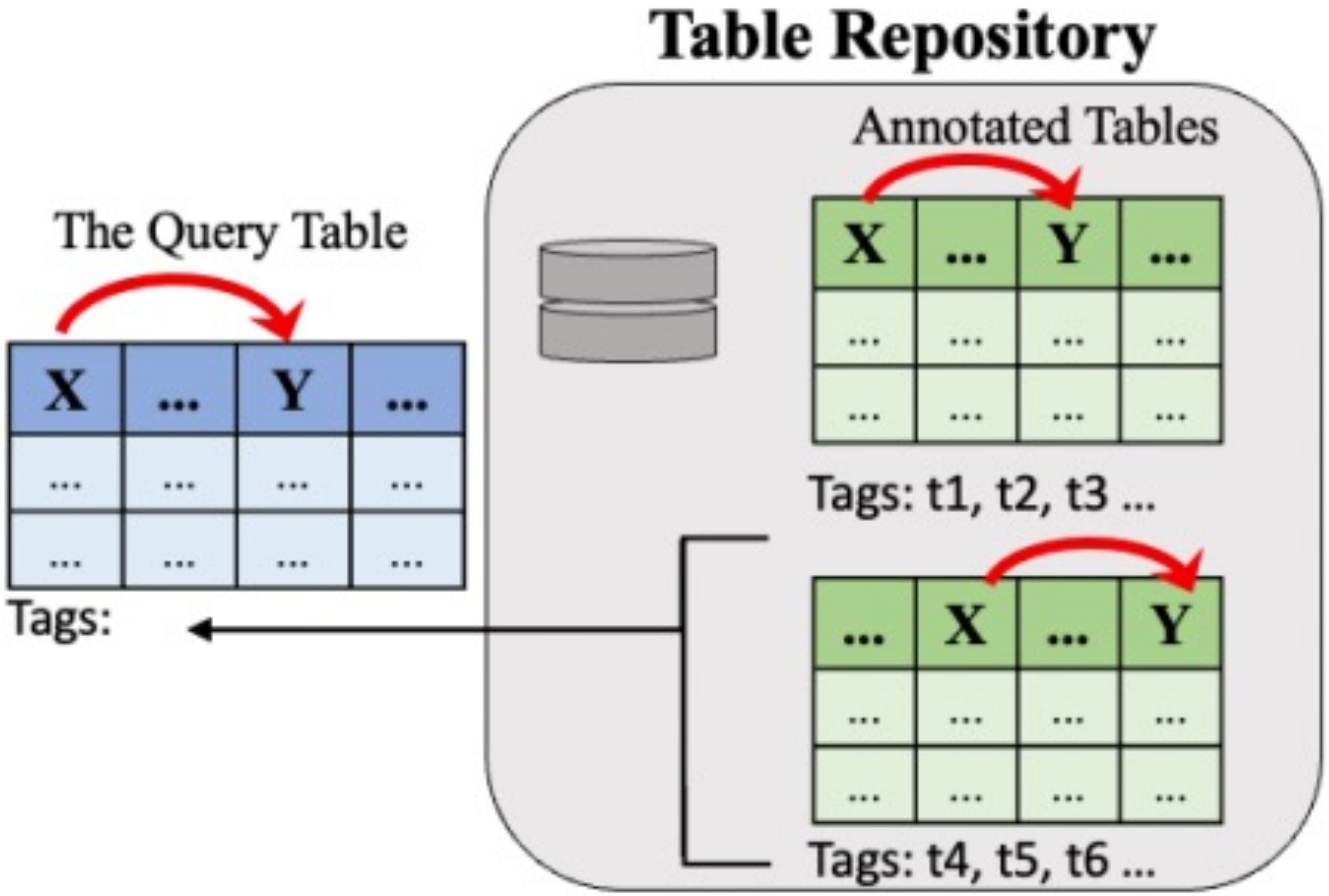
Details

Last Updated	May 14, 2021, 10:31 AM (UTC-07:00)
First Published Date	April 02, 2014
Maintainer	GIS Section
Data Update Frequency	Never

Background

Previous work on a tag recommendation algorithm

Still need human users to evaluate the recommendations



Objective

An interactive visual explainer that helps human users make tag augmentation decisions!

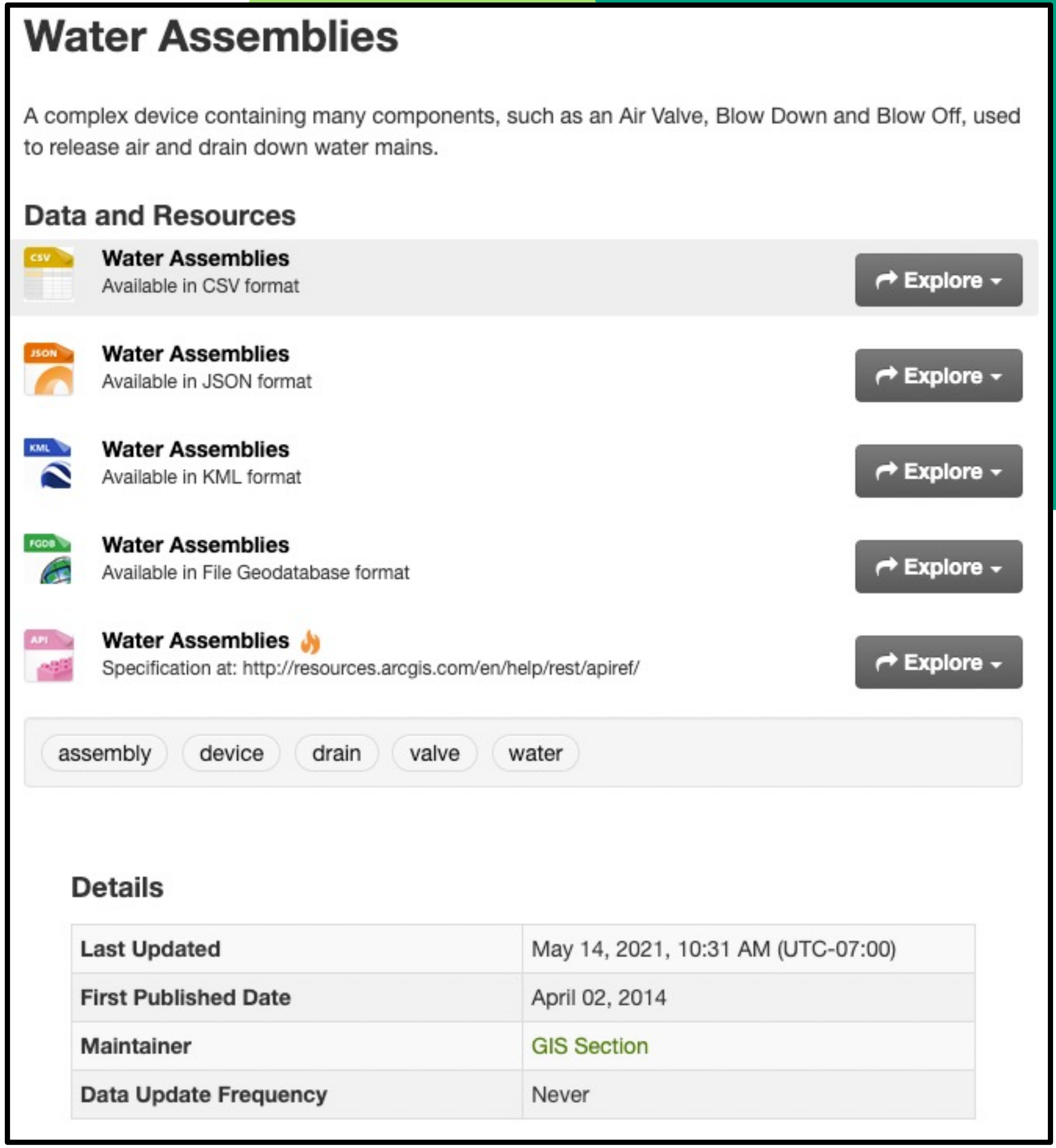
Data

160 tables from Surrey's Open Data site[1]

Every table is a single data item.

Explicit properties:

- Table name
- Column names
- Cell values
- Descriptive tags



Water Assemblies

A complex device containing many components, such as an Air Valve, Blow Down and Blow Off, used to release air and drain down water mains.

Data and Resources

Water Assemblies
Available in CSV format

Water Assemblies
Available in JSON format

Water Assemblies
Available in KML format

Water Assemblies
Available in File Geodatabase format

Water Assemblies
Specification at: http://resources.arcgis.com/en/help/rest/apiref/

assembly device drain valve water

Details

Last Updated	May 14, 2021, 10:31 AM (UTC-07:00)
First Published Date	April 02, 2014
Maintainer	GIS Section
Data Update Frequency	Never

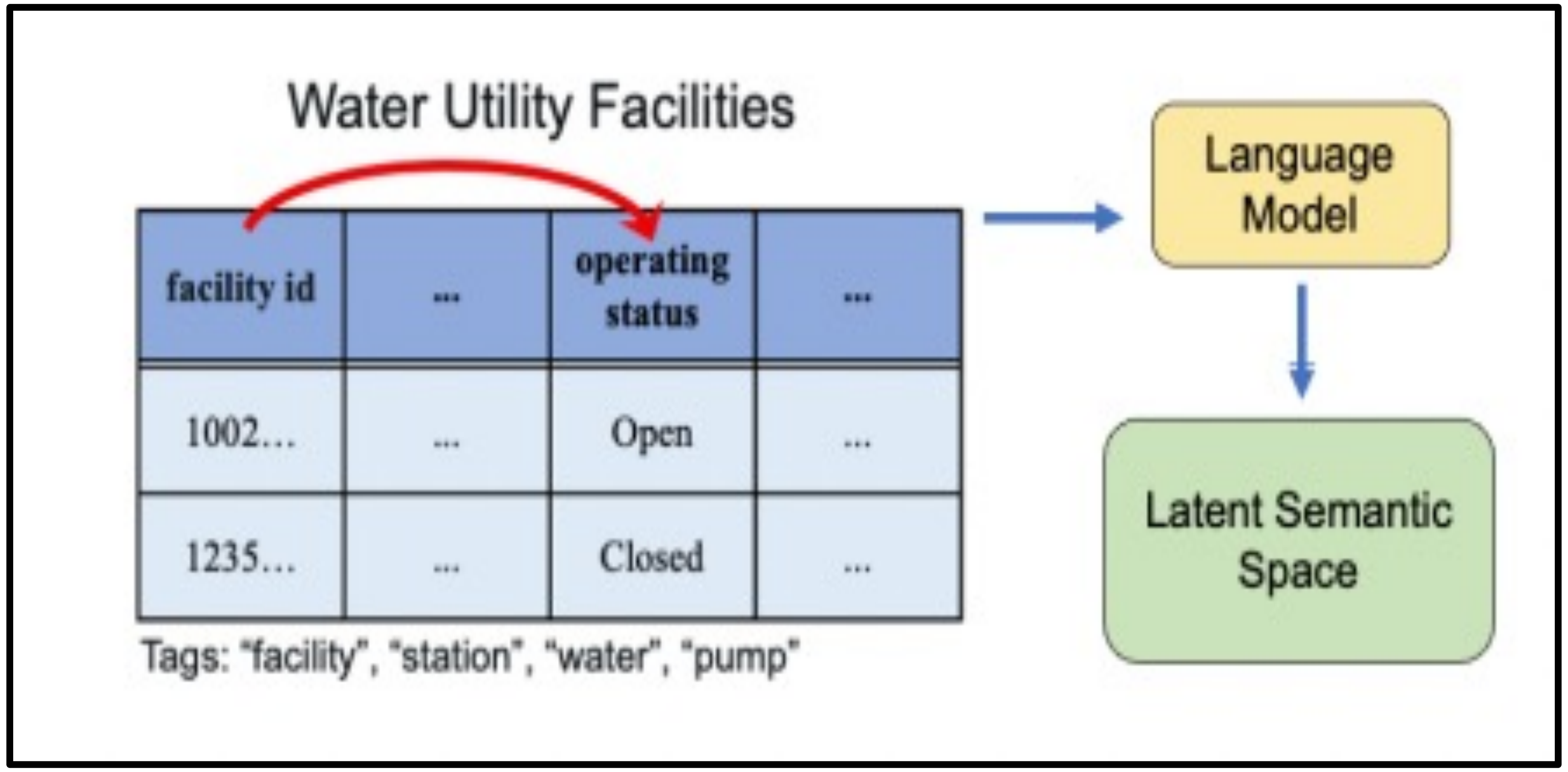
[1] https://data.surrey.ca

Data

160 tables from Surrey's Open Data site[1]

Implicit properties, used in the recommendation algorithm:

- intra-table correlated column pairs
  - » tables with the same column pair are related
  - » proof of table relatedness
- high-level semantic representations
  - » extracted by a language model
  - » exist in a high-dimensional space
  - » proof of table similarity



Water Utility Facilities

facility id	...	operating status	...
1002...	...	Open	...
1235...	...	Closed	...

Tags: "facility", "station", "water", "pump"

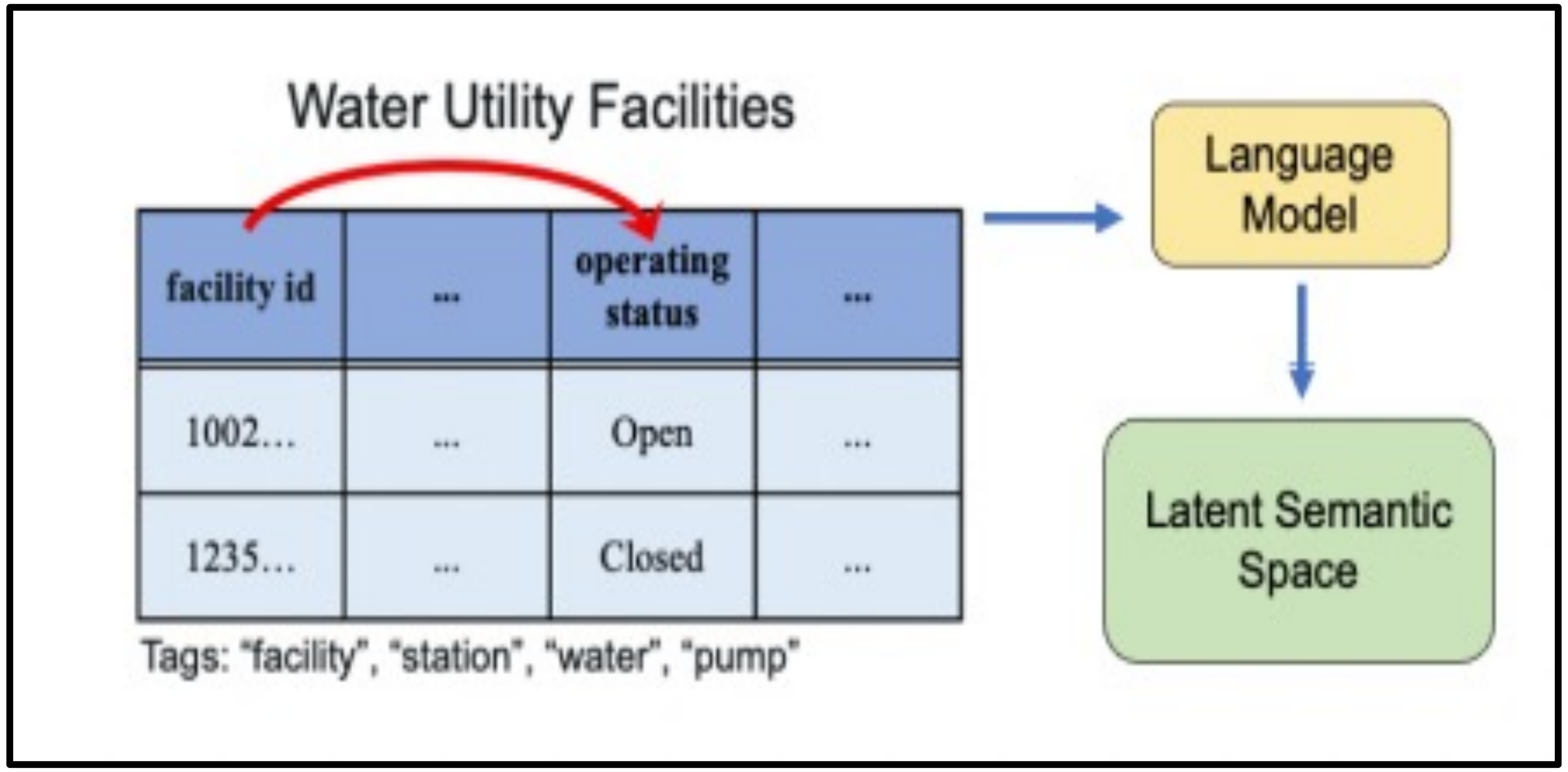
[1] https://data.surrey.ca

Data

160 tables from Surrey’s Open Data site[1]

Implicit properties, used in the recommendation algorithm:

- intra-table correlated column pairs
  - » tables with the same column pair are related
  - » proof of table relatedness
- high-level semantic representations
  - » extracted by a language model
  - » exist in a high-dimensional space
  - » proof of table similarity



Water Utility Facilities

facility id	...	operating status	...
1002...	...	Open	...
1235...	...	Closed	...

Tags: “facility”, “station”, “water”, “pump”

[1] https://data.surrey.ca

Tasks

Show the origin table of a recommended tag.

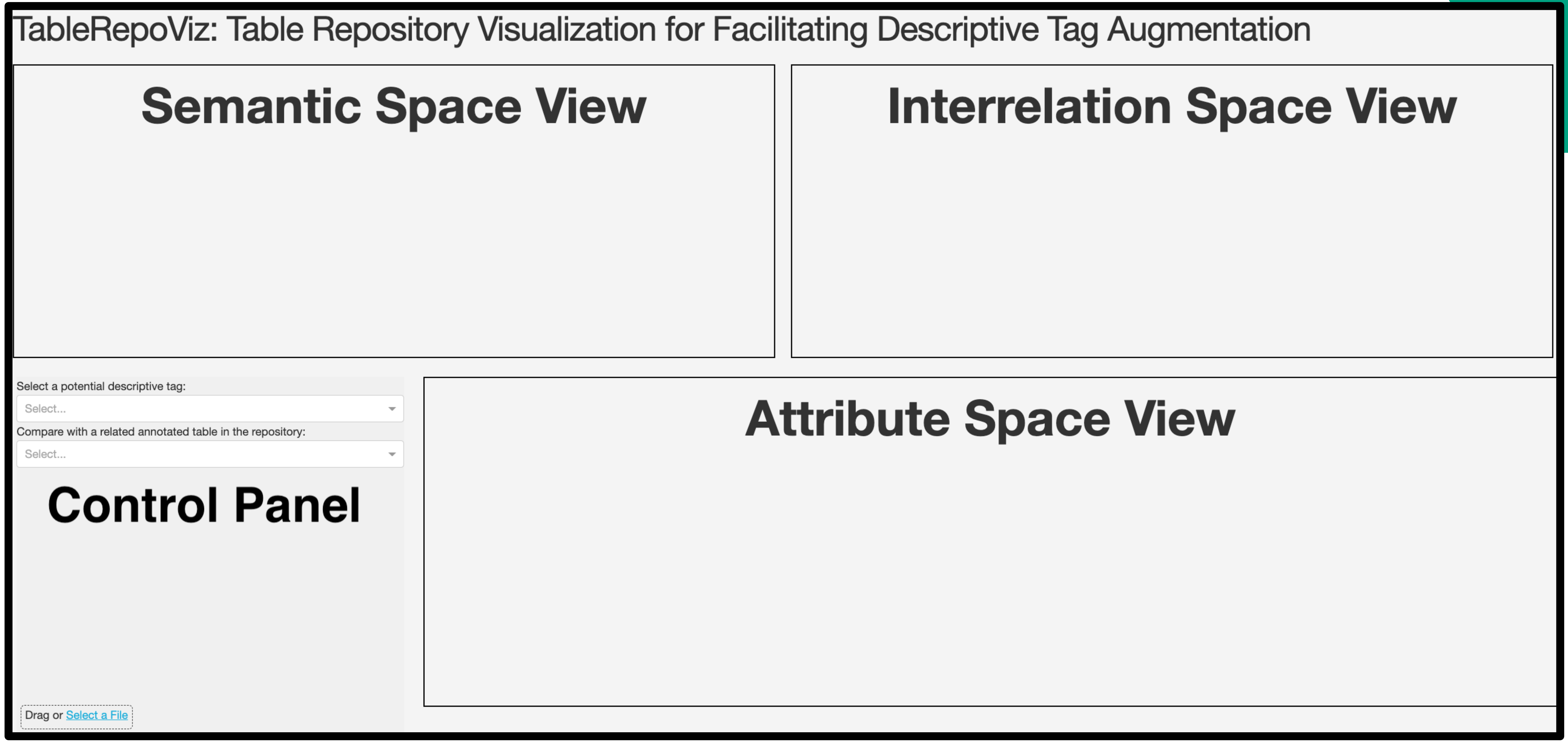
Explain why the algorithm recommends a tag.

Help to analyze the validity of a recommended tag.

Solution – Interface

TableRepoViz uses the notion of data spaces [1].

Each data space is dedicated to a different task.



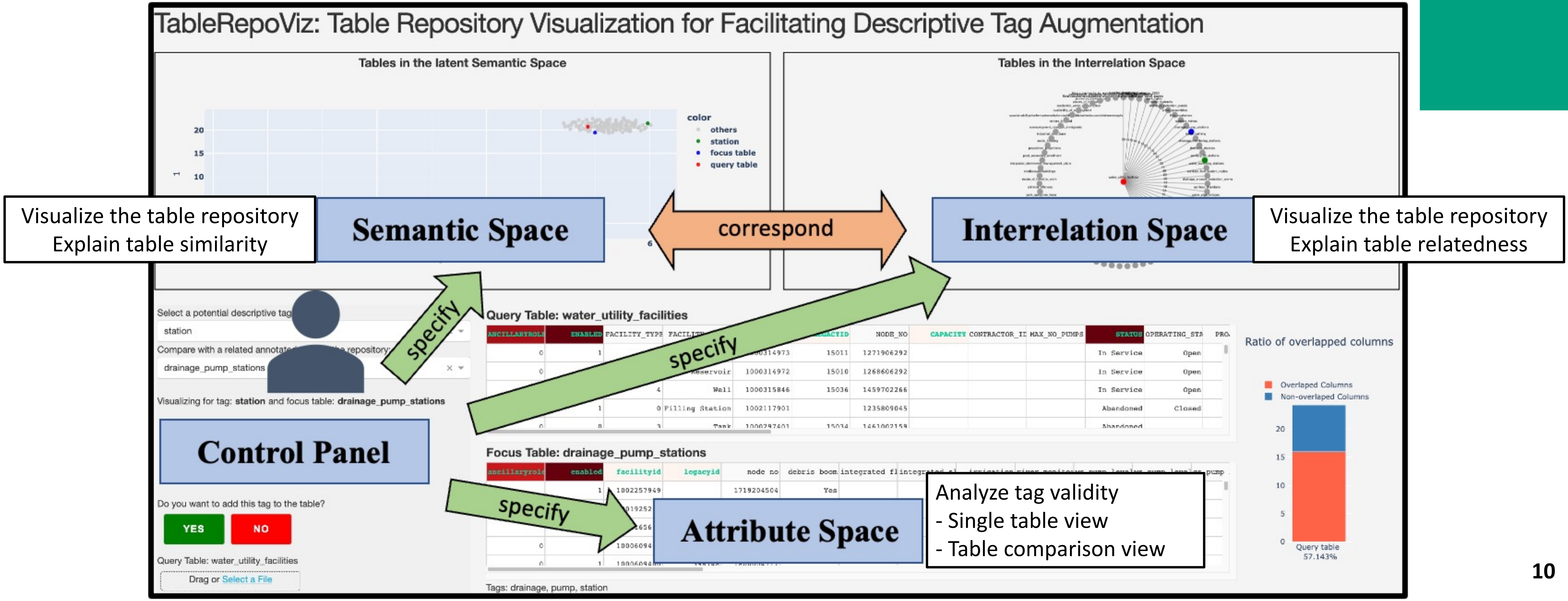
TableRepoViz: Table Repository Visualization for Facilitating Descriptive Tag Augmentation

[1] Z. Liu, Y. Wang, J. Bernard, and T. Munzner. Visualizing Graph Neural Networks With CorGIE: Corresponding a Graph to Its Embedding. IEEE Transactions on Visualization and Computer Graphics, 28(6):2500–2516, 2022.

Solution – Interface

Users can specify a focus tag and visualize it with the table repository.

User can select a table and compare it with the query table.



TableRepoViz: Table Repository Visualization for Facilitating Descriptive Tag Augmentation

Visualize the table repository
Explain table similarity

Semantic Space

correspond

Interrelation Space

Visualize the table repository
Explain table relatedness

specify

specify

specify

Control Panel

Attribute Space

Analyze tag validity
- Single table view
- Table comparison view

Select a potential descriptive tag: station

Compare with a related annotated table in the repository: drainage_pump_stations

Visualizing for tag: **station** and focus table: **drainage_pump_stations**

Do you want to add this tag to the table?

YES NO

Query Table: water_utility_facilities

Drag or Select a File

Query Table: water_utility_facilities

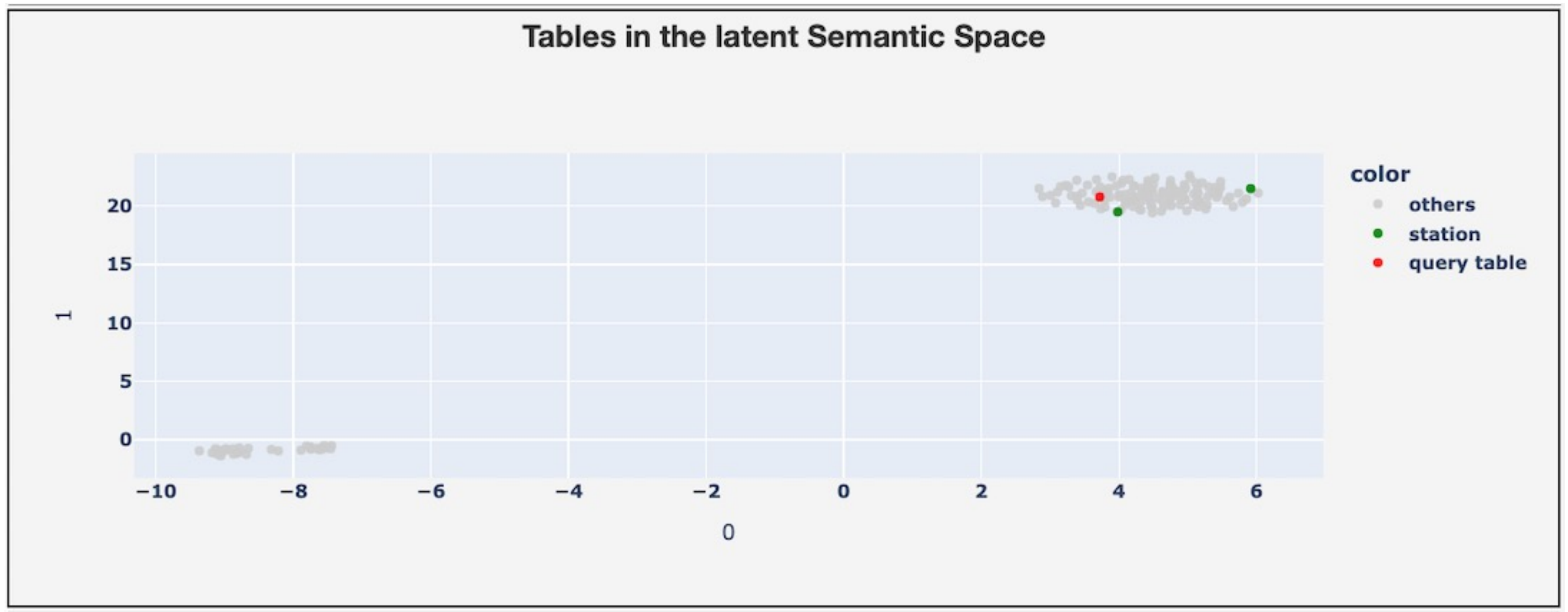
ANCILLARYROLE	ENABLED	FACILITY_TYPE	FACILIT	[illegible]	LEGACYID	NODE_NO	CAPACITY	CONTRACTOR_ID	MAX_NO_PUMPS	STATUS	OPERATING_STA	PRO
0	1	[illegible]	[illegible]	[illegible]00314973	15011	1271906292				In Service	Open	
0	[illegible]	[illegible]	Reservoir	1000314972	15010	1268606292				In Service	Open	
[illegible]	[illegible]	4	Well	1000315846	15036	1459702266				In Service	Open	
[illegible]	1	0	Filling Station	1002117901		1235809045				Abandoned	Closed	
0	0	3	Tank	1000297401	15034	1461002159				Abandoned		

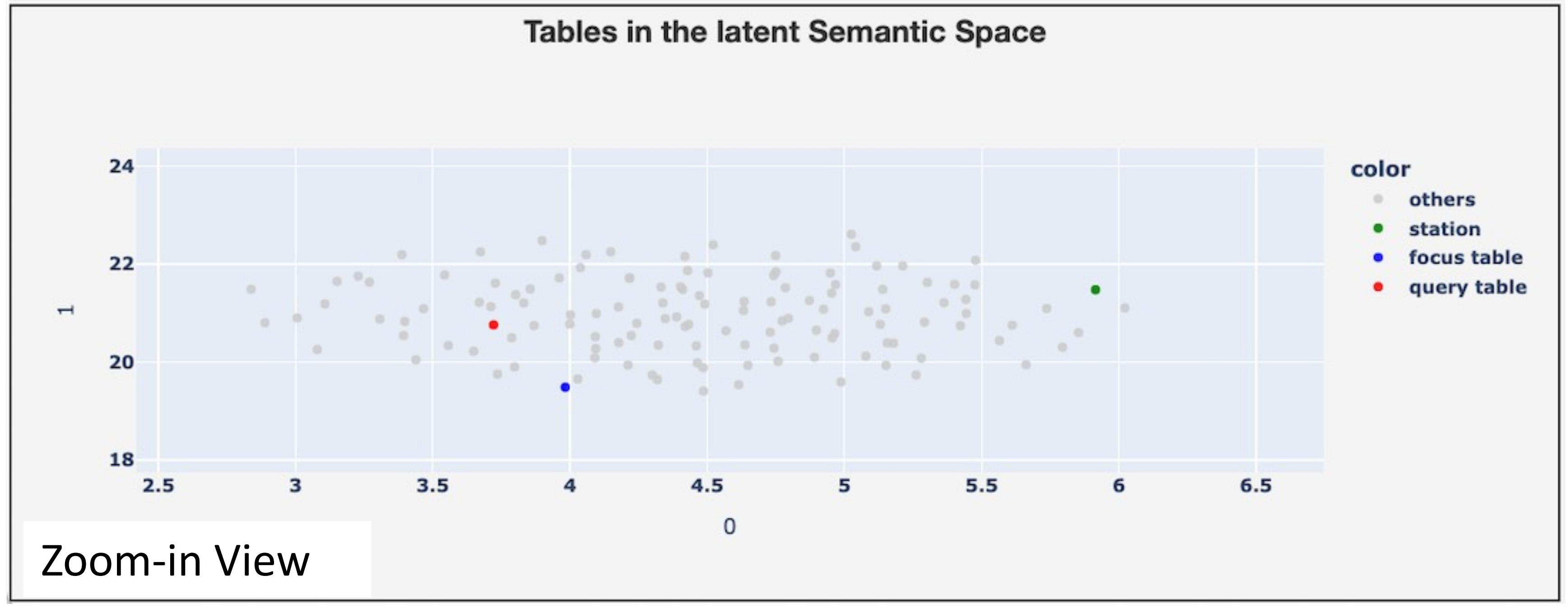
Focus Table: drainage_pump_stations

ancillaryrole	enabled	facilityid	legacyid	node no	debris boom	integrated fl	integrated sl	irrigation	river monitor	ws pump level	ws pump leve	gs pump
[illegible]	1	1002257949		1719204504	Yes							
[illegible]	[illegible]	[illegible]019252	[illegible]	[illegible]	[illegible]							
[illegible]	[illegible]	[illegible]656	[illegible]	[illegible]	[illegible]							
0	[illegible]	10006094	[illegible]	[illegible]	[illegible]							
0	1	10006094[illegible]	[illegible]	[illegible]	[illegible]							

Tags: drainage, pump, station

Solution – Semantic Space

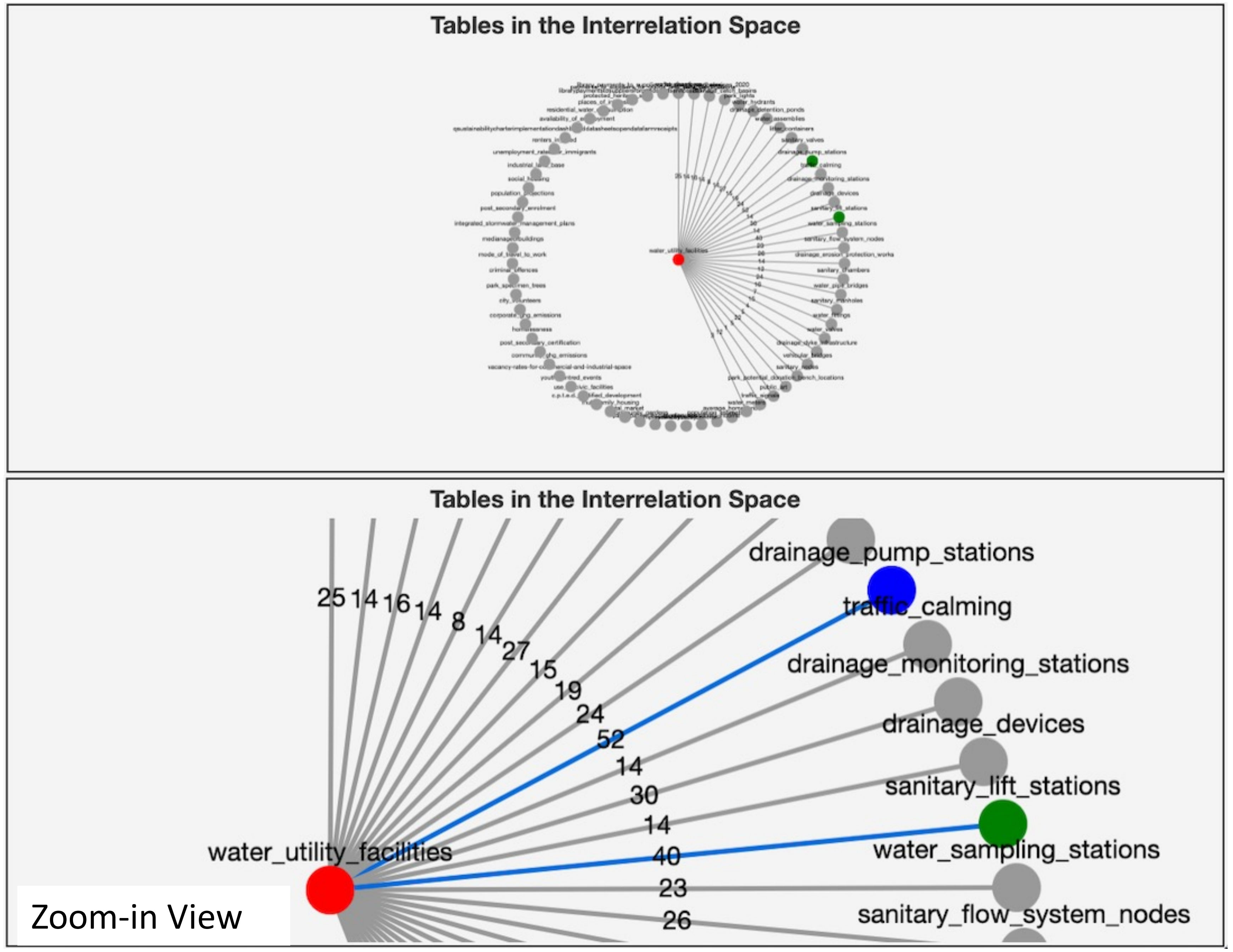

Tables in the latent Semantic Space


Tables in the latent Semantic Space

Zoom-in View

- Project table semantic representations into a 2D plane
- Encode table similarity with proximity
- Highlight the focus tag and user-selected table
- Explain language model usage in the recommendation algorithm

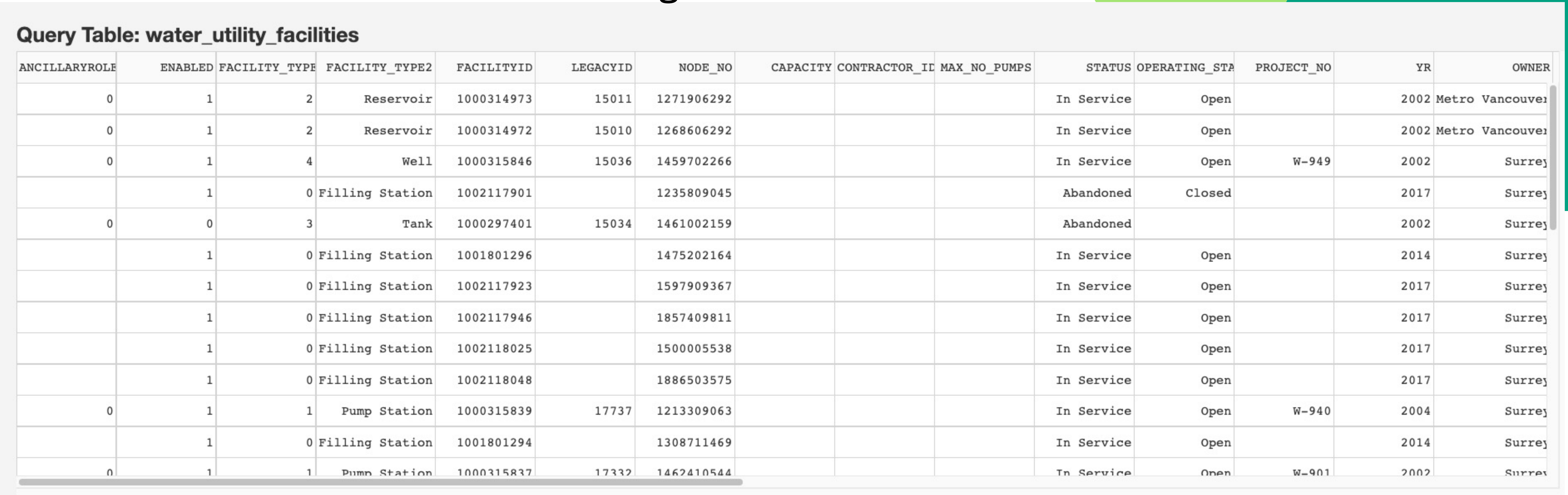
Solution – Interrelation Space



- Visualize the repository as a network to show implicit table connections
- Encode table relatedness with node connectivity
- Encode the number of shared correlated column pairs with edge weight
- Highlight the focus tag and user-selected table
- Explain the matching rules usage in the recommendation algorithm

Solution – Attribute Space

Single Table View



Query Table: water_utility_facilities

ANCILLARYROLE	ENABLED	FACILITY_TYPE	FACILITY_TYPE2	FACILITYID	LEGACYID	NODE_NO	CAPACITY	CONTRACTOR_ID	MAX_NO_PUMPS	STATUS	OPERATING_STA	PROJECT_NO	YR	OWNER
0	1	2	Reservoir	1000314973	15011	1271906292				In Service	Open		2002	Metro Vancouver
0	1	2	Reservoir	1000314972	15010	1268606292				In Service	Open		2002	Metro Vancouver
0	1	4	Well	1000315846	15036	1459702266				In Service	Open	W-949	2002	Surrey
	1	0	Filling Station	1002117901		1235809045				Abandoned	Closed		2017	Surrey
0	0	3	Tank	1000297401	15034	1461002159				Abandoned			2002	Surrey
	1	0	Filling Station	1001801296		1475202164				In Service	Open		2014	Surrey
	1	0	Filling Station	1002117923		1597909367				In Service	Open		2017	Surrey
	1	0	Filling Station	1002117946		1857409811				In Service	Open		2017	Surrey
	1	0	Filling Station	1002118025		1500005538				In Service	Open		2017	Surrey
	1	0	Filling Station	1002118048		1886503575				In Service	Open		2017	Surrey
0	1	1	Pump Station	1000315839	17737	1213309063				In Service	Open	W-940	2004	Surrey
	1	0	Filling Station	1001801294		1308711469				In Service	Open		2014	Surrey
0	1	1	Pump Station	1000315837	17332	1462410544				In Service	Open	W-901	2002	Surrey

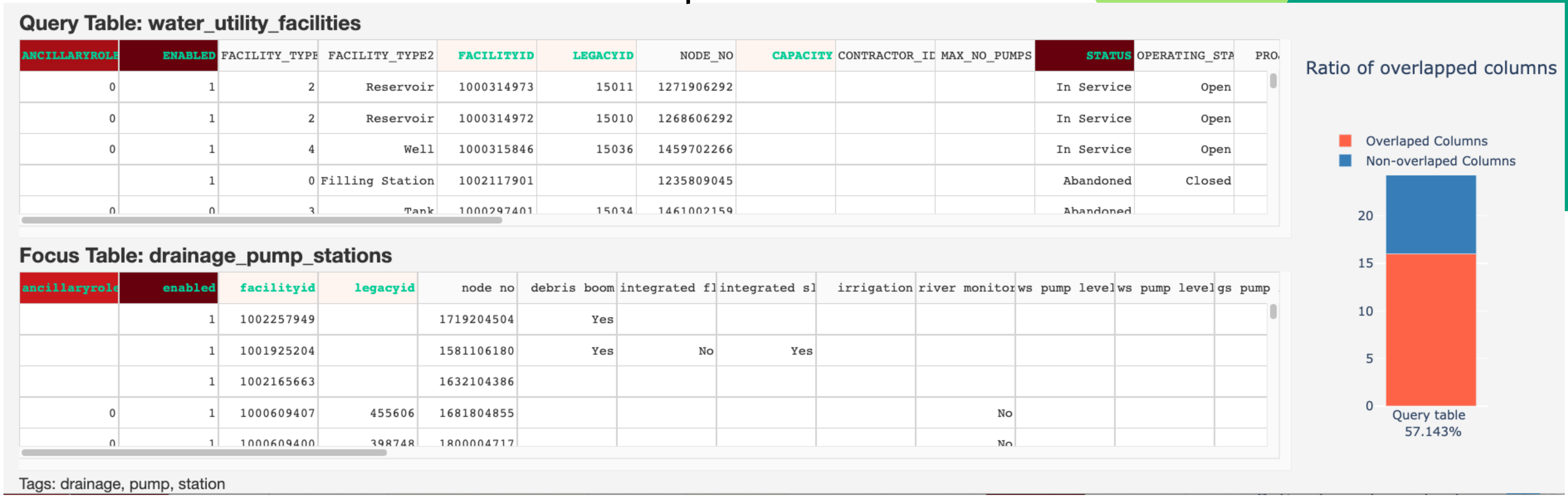
- Present the cell values in the query table.

- Users can explore data in an undirected manner.

- Verify if the tag is accurate in describing the query table.

Solution – Attribute Space

Comparison View



Query Table: water_utility_facilities

ANCILLARYROLE	ENABLED	FACILITY_TYPE	FACILITY_TYPE2	FACILITYID	LEGACYID	NODE_NO	CAPACITY	CONTRACTOR_ID	MAX_NO_PUMPS	STATUS	OPERATING_STA	PRO
0	1	2	Reservoir	1000314973	15011	1271906292				In Service	Open	
0	1	2	Reservoir	1000314972	15010	1268606292				In Service	Open	
0	1	4	Well	1000315846	15036	1459702266				In Service	Open	
	1	0	Filling Station	1002117901		1235809045				Abandoned	Closed	
0	0	3	Tank	1000297401	15034	1461002159				Abandoned		

Focus Table: drainage_pump_stations

ancillaryrole	enabled	facilityid	legacyid	node no	debris boom	integrated fl	integrated sl	irrigation	river monitor	ws pump level	ws pump level	gs pump
	1	1002257949		1719204504	Yes							
	1	1001925204		1581106180	Yes	No	Yes					
	1	1002165663		1632104386								
0	1	1000609407	455606	1681804855					No			
0	1	1000609400	398748	1800004717					No			

Tags: drainage, pump, station

- A stacked bar chart to illustrate the ratio of overlapped column names in the query table.

- Heatmap-like column headers to visualize the percentage of shared values in the overlapped columns.

- Verify if applying the tag to the query table is consistent with past annotations.

Limitations & Future Work

Lack of interactivity in data spaces

- Cannot directly select a table in data spaces.
- Support visualization correspondence across all three data spaces.

Limited to column-wise inference

- Caused by the underlying recommendation algorithm
- Explore other visualization idioms for comparing tables

Scalability issues

- Does not scale for large tables, huge computation costs rendering the table comparison view.
- Does not scale for a large repository, may cause clutter in data space views.
- Try data reduction idioms to visualize part of the data.

Questions?