Data Representations Discussion

April 25, 2024

Goal

As a group, try to decide: which of these categories of data representations should NC-BPAID pursue schema development in?

Summary

Following an overview and examples of data representation categories, participants discussed the following questions:

1. How would you formally define each category of data representation?
2. What are the drawbacks and benefits of each representation type?
3. Are there any other ways of representing these data that are not covered by these categories?
4. Which of these representations are important to standardize? Which comes first?
5. What are the most important use cases satisfied by each of these representations?

Data Representation Categories

> *Note: The working terminology we use here is meant to describe the dimensionality of information about - for example - a sidewalk. These terms are not intended to correspond to the way dimensionality is used in other fields. If you have ideas about how we can make these representation categories clearer, please comment.*

1-Dimensional

The actual location of the infrastructure is not mapped: there's no geospatial feature that describes the real world location of infrastructure. The presence of the infrastructure and its characteristics are encoded in attributes and attribute values of a different dataset, like a roadway centerline.

Some participants perceived that this category underestimates the power of Linear Referencing Systems (LRS). The benefits of this representation include the smaller amount of storage space required, the existing widespread use of LRS (especially by state DOTs), the ability to condense lots of information to a centerline, and the utility for asset management and data-driven safety analysis. The drawback of this representation is that it is difficult or unclear how to move between this representation and a routable graph necessary to support trip planning/navigation and network analysis.

2-Dimensional

The actual location of infrastructure is mapped: there is some kind of geospatial feature describing the real world location of the infrastructure.

Disconnected Linear Features

The geospatial feature describing the real world location of the infrastructure is linear, and the linear features are not connected to one another, even if the infrastructure in the real world is connected or continuous. There are attributes attached to these linear features.

Participants viewed this as the category being most widely used at this time. The benefits of this representation are that it is conceptually simple and can be used for asset inventory. The drawback of this representation is that it does not capture the connectivity between features needed to support trip planning/navigation and network analysis.

Connected Linear Features

The geospatial feature describing the real world location is represented by nodes connected by links, to create a network (also known as a graph). There are attributes, or tags, attached to both links and nodes.

Many participants had a bias towards connected linear features being the most broadly useful data representation. The benefits of this representation include that the topological relationship between features is explicit, its coverage of use cases (including inventory and data-driven safety analysis, but especially trip planning/navigation and network analysis), the ability (but not requirement) to provide a high level of detail, and modest storage space requirements. Its drawbacks are that it is difficult to move between this representation and an LRS.

Polygonal Features

The geospatial feature describing the real world location of the infrastructure is represented by individual polygon features with attributes attached to the polygons.

The benefits of this representation are that it is useful for asset management/maintenance, it is possible to incorporate polygons into a network/graph for routing, and it’s easy to derive important attributes like width and setback, which influence user experience. Drawbacks are that the inclusion of polygons into a network/graph increases complexity and makes it more difficult to interpret.

3-Dimensional

A 3D digital model of the infrastructure, tied to GPS location. It's typical, but not required, that geospatial features (points, lines, polygons) and attributes are derived from this 3D digital model.

Participants generally agreed that this representation provided more information than necessary for most use cases, but imagined that 3D data, combined with imagery would, in the future, become the raw data collected and stored by data producers, from which data in any other representation could be derived based on specific needs which may bypass the need to focus

on a single data representation. Benefits of this representation include their use in digital twins, and coverage of use cases like ADA transition planning, where a high level of detail is required. Drawbacks include the complexity and expense of data storage, the difficulty of working with and managing 3D data, and the feasibility of scaling 3D data. Participants also noted that there are different levels of granularity within 3D data that may be simpler to use: for instance, a digital elevation model can provide useful attributes (e.g., grade) and also scale (e.g., USGS 3DEP).

Gaps

- Several participants disagreed that the categories described above are appropriate categories and definitions of the different ways to represent the data of interest. Participants suggested using already formalized categories from resources like the Building Smart International Overall Architecture Guidelines or defining different categories ourselves that align more with the use cases.
- Several participants raised that there are differences in the naming conventions (e.g., separated bike lane, protected bike lane, cycle track) as well as the definitions of the infrastructure (e.g., are crossings over driveways and alleys sidewalks or unmarked crossings?). It would be useful to draw on existing resources like NACTO and the MUTCD to clarify the names and definitions of the infrastructure types we are interested in describing.
- Several participants noted the importance of considering how the data representations connect to the data of other modes (e.g., GTFS) or integrating/joining other key data (e.g., trees → shade cover) , which was not part of this discussion.
- Several participants identified the connection between our data of interest and ADA/PROWAG and MIRE.
- Several participants identified that it would be useful to create a hierarchy of representations and the use cases they are suitable for and tools or practices for converting data between representations.

Key Findings

- Sentiment seemed to favor the connected linear features representation due to the necessity of a connected network/graph to fulfill the key use case of routing/trip planning (identified in NC-BPAID's purpose), and other use cases: conducting network analysis, data-driven safety analysis, and asset inventory.
- One topic that NC-BPAID should explore further is the relationship between data made up of connected linear features and Linear Referencing Systems (LRS).
- The reality that data already exists and will continue to be produced for the foreseeable future in all of these representations requires continued clarification and communication of the data categories, their benefits and drawbacks, the use cases they satisfy, and development of tools, methods, and/or practices for moving between them.