# The Hitchhiker's Guide to Analyzing the FCC Broadband Data Collection Datasets

Jonatas Marques, Alexis Schrubbe, Nicole P. Marwell, Nick Feamster
University of Chicago

## Abstract

The FCC Broadband Data Collection (BDC) program has had–and will continue to have–tremendous impact on directing policy interventions and funding towards the goal of achieving broadband equity, access, and deployment across the United States. In this paper, we share our experience analyzing the data disseminated by the FCC as part of this program. We focus on discussing the challenges and limitations that one may encounter when exploring the datasets made publicly available as part of this program. Examples are the lack of direct, public data on the fabric layer; the retroactive removal of availability records from past data releases; and the purely file-based data serving model. We provide recommendations to stakeholders on ways to overcome these challenges and cope with limitations. These recommendations seek to introduce best practices for processing and analyzing the BDC data. Where appropriate, we also bring suggestions to the FCC on approaches to eliminate data limitations and lower barriers to analysis. These suggestions involve changes to how BDC data is published, served, updated, and summarized by the FCC.

## 1 Introduction

In November 2021, the Broadband Equity, Access, and Deployment (BEAD) program allocated 42 billion dollars to expand high-speed Internet access across the United States (US) [10, 16]. As part of this initiative, the US Congress tasked the Federal Communications Commission (FCC) with developing a national map on broadband deployment and availability, the National Broadband Map (NBM) [3, 4, 15]. This map was the key determinant to direct BEAD investments to areas in need of broadband infrastructure improvements at the federal stage. The FCC encouraged public participation in refining the NBM through the submission of "challenges" to either locations on the map or the status of availability at these locations [9]. These challenges allowed citizens and organizations to report discrepancies between the map's data and actual broadband availability, with the goal of ensuring a more effective distribution of funds. Given the dependence that BEAD (and possibly future investments of the kind) has on the NBM, it is of the utmost importance to analyze the data contained in it, its accuracy, and its evolution. The FCC has regularly published data on the NBM, which is collected through the Broadband Data Collection (BDC) program [2]. The BDC datasets represent the main pathway for researchers and policymakers to assess whether the BEAD program is moving towards its goals and for learning lessons for future investments of this kind.

In the context of our research group, we set out to analyze the BDC dataset seeking to answer several research questions related to the BEAD challenge process, such as: Who were the winners (and the losers) in the challenge process? How did different communities engage with the challenge process?; Were there any social or demographic factors that influenced this engagement? Does this challenge process reinforce pre-existing inequalities impacting historically disadvantaged communities? As part of our effort, we have directly mentored several undergraduate and graduate students in Data Science, Computational Analysis and Public Policy participating in quarter-long data clinic courses between September 2023 and May 2024. In these courses, students have the opportunity to practice their learned skills with the goal of building data pipelines that facilitate the analysis of the BDC data.

In this paper, we share our experience working with the BDC dataset which was an adventure filled with—no pun intended—challenges. Since we have set out on this journey, our intention is to invite hitchhikers along so that our efforts can construct a helpful research-based guide for future explorers interested in the BDC dataset. Our main contributions are summarized as follows.

- Identify the main challenges and limitations one may encounter when exploring and analyzing the datasets made publicly available as part of the FCC Broadband Data Collection program.
- Advise practitioners, researchers, policymakers, and the academic community on approaches to overcoming

the identified challenges and cope with the currently existing limitations of the datasets.

- Suggest changes to the BDC data publication, serving, update, and summarization processes that could be implemented by the FCC to facilitate engagement with the datasets by all interested parties.

This remainder of this paper is organized in two main sections. In Section 2, we provide an overview of the Broadband Data Collection (BDC) program datasets. In Section 3, we introduce and discuss the challenges and limitations we observed during our analysis of these datasets.

# 2 The FCC Broadband Data Collection (BDC) Datasets

In this section, we briefly describe the datasets published by FCC as part of the Broadband Data Collection (BDC) program and available for download through the National Broadband Map portal (Figure 1) [3]. These datasets comprise three main sources of data: reported broadband availability, challenges on reported availability, and challenges over the broadband-serviceable fabric. The BDC portal offers most data downloads formatted as compressed comma-separated values (CSV) files [13]. For some mobile availability data, the portal offers downloads as both compressed ESRI Shapefile [1] and compressed GeoPackage files [11]. The portal presents a few pre-defined views to the data from each source, as we describe in the following sections[1].

## 2.1 Availability Dataset

As the name suggests, the availability dataset contains data that indicates if and what type of broadband Internet access is *available* across the United States and its territories. The Broadband Data Collection (BDC) program captures two major types of Internet access: *fixed broadband* and *mobile broadband*. Depending on the type of access, availability data is published in slightly different ways. The main distinction lies in that *mobile* access availability is reported for geographical areas whereas *fixed* broadband availability is reported for individual locations. Regardless of the type of access, availability data is offered in raw form as well as in multiple summary forms.

### 2.1.1 Raw Data

*Fixed broadband availability* is reported for every known broadband-serviceable location (BSL) in the nation. A BSL is generally defined as a location at which mass-market fixed

[1]A list of FCC-registered providers is also available through the portal.

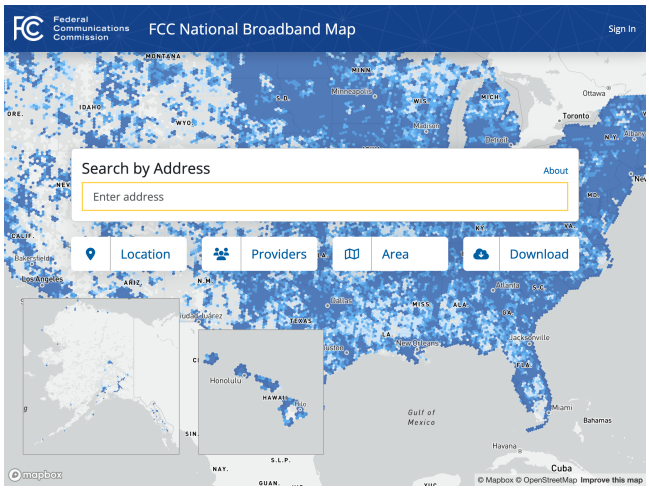


Figure 1: Home page of the National Broadband Map portal with options to lookup availability by location, providers, and area as well as to access the Broadband Data Collection downloads.

broadband Internet access service is, or can be, installed. A BSL may be considered residential (e.g., single-family homes, duplexes, apartment buildings), business (e.g., restaurants, dry cleaners, day cares, clothing stores), or both (i.e., mixed-use buildings) [7]. The BDC dataset for fixed availability holds zero or more records (or table rows) for each known BSL. Each row represents an Internet subscription offer by a single provider via a specific access technology for that BSL. Currently, the FCC recognizes the following access technologies for fixed Internet access: Copper, Cable, Fiber to the Premises, Geostationary Satellite, Non-geostationary Satellite, Unlicensed Fixed Wireless, Licensed Fixed Wireless, and Licensed-by-Rule Fixed Wireless. In addition to fields identifying the BSL, provider, and access technology, each row also indicates the maximum download and upload speeds advertised by the provider as well as whether the latency at the location is expected to be "low" (i.e., under 100 milliseconds). For data download purposes, the raw fixed availability is always first split by states[2] (see Figures 2 and 3). For any given state, raw data is then split by access technology (right panel in Figure 2) or provider (left panel in Figure 3). In other words, any single downloaded file will contain data either (a) for a single state and a single technology (and most likely multiple providers) or (b) for a single state and a single provider (and possibly multiple technologies). More details on the raw fixed broadband availability data downloads can be found in

[2]Throughout this paper we use the term states to refer to both states and territories of the United States.

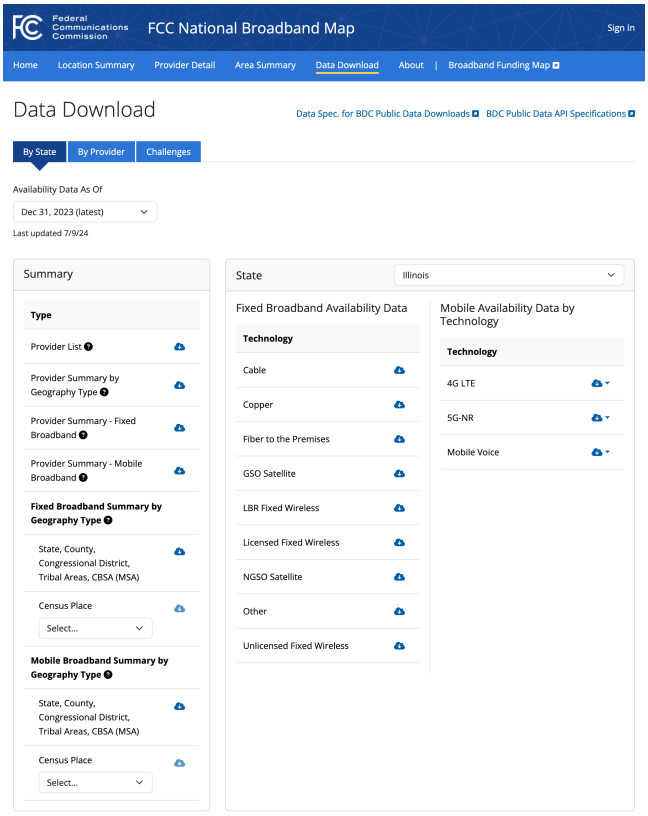


Figure 2: Webpage for downloading availability data files. The right panel provides access to summary data and the left panel to the raw data.

Sections 3.1.1 and 3.2.1 of the FCC's data specification document [5].

In its raw form, *mobile broadband availability* is reported for pre-defined geographical areas. The FCC uses the H3 geospatial indexing system [14]—which partitions the world into hexagonal cells—to determine unique areas across the nation for which providers can report availability. In its current version, the dataset uses H3 resolution 9 cells, which are on average about 0.105 square kilometers (or 0.041 square miles) in area (see Figure 4). This amounts to about 4.8 billion individual cells across the globe. The dataset holds zero or more records (or table rows) for each H3 cell. Each record (or table row) in this data represents the presence of a wireless signal of a specific mobile access technology from a specific provider in a specific cell. Apart from fields identifying the cell, provider, and technology, each record also indicates (a) the minimum expected download and upload speeds in the area, (b) the minimum expected signal strength (expressed in decibels) in the area, and (c) whether the area is modeled to have coverage only when user equipment is in outdoor stationary environments or also within in-vehicle mobile environments. For download purposes, the raw mobile availability data is always split by state, provider, and access technology (e.g., 3G, 4G, 5G) and served as ESRI Shapefiles or GeoPackage files (left panel in Figure 3). In other words, any single file will contain data for a single state, a single provider, and a single wireless technology. More details on the raw mobile availability data downloads can be found in Section 3.2.2 of the FCC's data specification document [5].

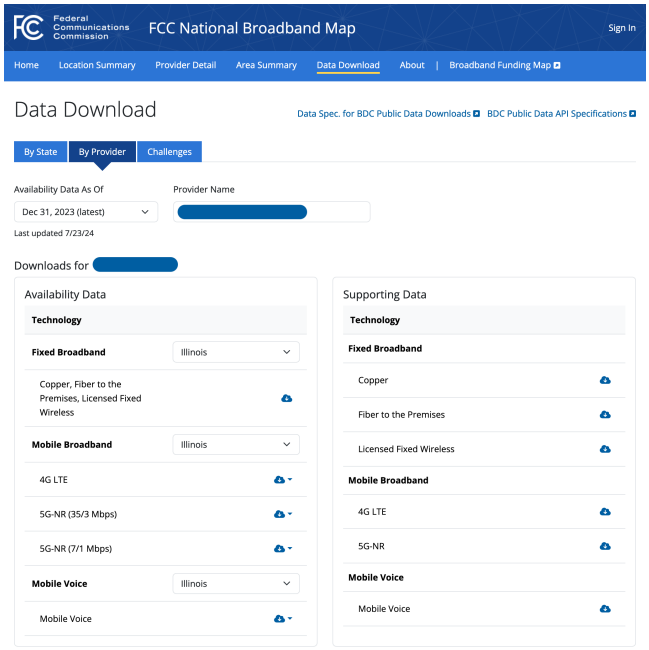


Figure 3: Webpage for downloading availability data files for a specific provider. The left panel gives access to the raw data and the right panel to the provider's supporting documents.

### 2.1.2 Summary Data

The BDC portal offers six different options for availability data summarization (considering both fixed and mobile broadband access), as we describe in the next paragraphs. More details on the availability summary data downloads can be found in Sections 3.1.2 and 3.1.3 of the FCC's data specification document [5].

**Provider Summary by Geography Type.** This download option consists of a single compressed CSV file (left panel in Figure 2) where both fixed and mobile broadband coverage is summarized for each individual provider across multiple geographical units. These geographical units include states, counties, congressional districts, census places, tribal areas, and core-based statistical areas (CBSAs). Besides fields identifying the geographical unit (e.g., its type, FIPS code, and

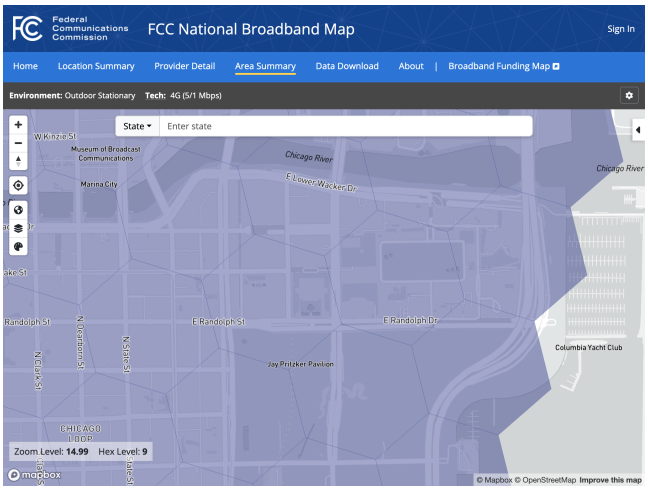


Figure 4: Visualizing mobile Internet access availability within H3 resolution 9 hexagon cells in the National Broadband Map website.

description) and the provider, each record indicates the type of data being summarized (either fixed or mobile broadband) and two percentage fields of varying meaning. Whenever a record summarizes fixed broadband availability, the first (second) percentage field indicates the percentage of BSLs contained within the geographical unit for which the provider reports residential (business) service (respectively). Whenever a record summarizes mobile access, the first (second) percentage field indicates the percentage of area within the geographical unit for which the provider reports coverage for outdoor stationary environments (for in-vehicle mobile environments, respectively).

**Provider Summary - Fixed Broadband.** This option consists of a single compressed CSV file (left panel in Figure 2) where fixed broadband availability is summarized for each individual provider across multiple *fixed* broadband access technologies. Records in this file include fields to identify the provider and the technology as well as four summary fields indicating the total count of BSLs buildings and individual units for which the provider reports residential or business fixed broadband service with the identified technology.

**Provider Summary - Mobile Broadband.** This option is very similar to the previous except that is summarizes mobile availability for each individual provider across multiple *mobile* broadband access technologies. Records in this file include fields to identify the provider and the technology as well as two summary fields indicating the total area (in square kilometers) for which the provider reports mobile broadband service only in outdoor stationary environments or inside in-vehicle mobile environments.

**Fixed Broadband Summary by Geography Type.** This download option summarizes *fixed* broadband coverage for each individual geographical unit across multiple geographical levels and access technologies, regardless of the specific providers offering service. The geographical levels include states, counties, congressional districts, census places, tribal areas, and core-based statistical areas (CBSAs). This data may be offered as a single compressed CSV file containing units from all geographical levels or, in more recent releases, multiple files, each containing data for units from groups of geographical levels (left panel in Figure 2). Records in these files contain fields identifying and describing the geographical unit and access technology, counting the total number of BSL units in the geographical unit, indicating what kind of service is available at the location (i.e., residential-only, business-only, or both), and multiple summary fields. Each summary field indicates the percentage of BSL units within the geographical unit for which at least one provider reports fixed broadband service with speeds at or above multiple pre-defined threshold pairs (in Mbps) for download and upload, such as 25 Mbps for download and 3 Mbps for upload, or 100/20 Mbps, or others.

**Mobile Broadband Summary by Geography Type.** This option summarizes *mobile* broadband coverage for individual geographical units of multiple geographical levels (same levels described in the previous paragraph), regardless of the specific providers offering service. This data may be offered as a single compressed CSV file containing units from all geographical levels or, in more recent releases, multiple files, each containing data for units from groups of geographical levels (left panel in Figure 2). Records in these files contain fields identifying and describing the geographical unit as well as a total of eight summary fields. describing coverage across access technologies (i.e., 3G, 4G, and 5G), speeds (for 5G, either 7/1 Mbps or 35/3 Mbps), and service model (i.e., outdoor stationary-only or in-vehicle mobile environment). The summary fields represent the percentage of area within the geographical unit for which any provider reports mobile broadband service using a particular access technology (i.e., 3G, 4G, and 5G), for user equipment either at outdoor stationary-only or at in-vehicle mobile environments, and—in the case of 5G— with speeds at or above pre-defined thresholds.

**Mobile Availability Data by Technology.** This last download option summarizes *mobile* broadband availability for each individual H3 resolution 9 cell across access technologies, regardless of provider. This data is split into multiple compressed ESRI Shapefile or GeoPackage files by state and access technology. Each file contains summary data for a single state and technology. In addition to fields identifying the H3 cell and the access technology, each record also indi-

cates the minimum download and upload speeds expected in the area (from any provider) and whether the area is modeled to have coverage only when user equipment is in an outdoor stationary environment or also in in-vehicle mobile environments.

To conclude this section, we note that the BDC portal also enables downloading files containing supporting data submitted by Internet providers to the FCC. These files contain methodological information on the approaches (e.g., signal propagation models, cross-referencing of customer addresses and BSL addresses) applied by providers to determine where their service is available along with justification for using their chosen approaches.

## 2.2 Availability Challenge Dataset

In order to continually improve the accuracy of the National Broadband Map (NBM), the FCC allows for and encourages challenges to be raised by citizens and organizations whenever they observe discrepancies between the data in the map and actual broadband availability, as per their lived experience. Challenges may be submitted for reasons such as the access technology or advertised speeds not actually being offered at a location, the provider having denied a request for service at the location (despite advertisement, for example), and a wireless or satellite signal not being available at the location (despite propagation models suggesting it would be available, for example). The full list of reasons for submitting a challenge can be found in Reference [5]. All raised challenges are adjudicated by the FCC, usually requesting additional supporting evidence from the implicated provider to counter the challenge. Whenever a challenge is upheld, changes are made to the BDC dataset to reflect the updated map.

As part of this BDC challenge process, the FCC regularly publishes data on the submitted challenges and their outcomes (see Figure 5). This constitutes the availability challenge dataset. New data on availability challenges are released monthly and offered in two forms: raw and cumulative summary. The raw challenge data is offered for both fixed and mobile challenges. This data is split into CSV files by broadband type (fixed or mobile), state, adjudication status, which can be either in progress or resolved. Each record in the raw data contains information on a single challenge. The core fields in these records—regardless of adjudication status—are a unique identifier for the challenge, the identifier for the BSL (or H3 cell in case of mobile broadband), the state where the BSL (or cell) is located, the unique identifier of the provider being challenged, the access technology subject to challenge, the reason for the challenge submission, and the date the challenge was submitted. For resolved challenges, each record also includes the outcome of the challenge (whether it was upheld,

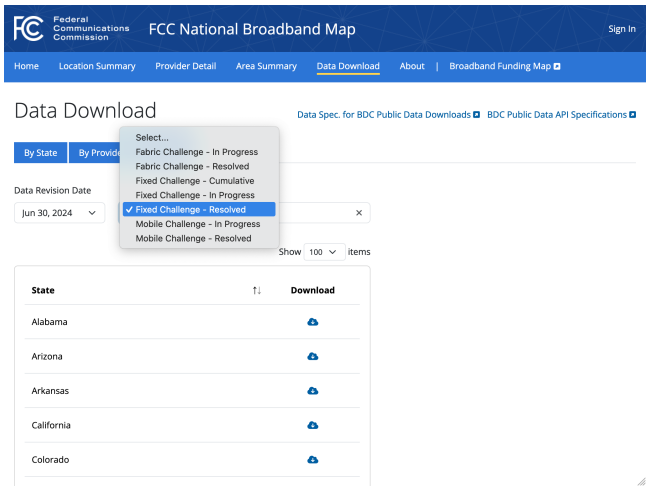


Figure 5: Webpage for downloading availability and fabric challenge data files.

overturned, or withdrawn), how the challenge got to such outcome (whether by concession by the provider, changes in service, or adjudication by the FCC), and the date when the challenge was adjudicated. The cumulative summary data is only available for fixed broadband availability challenges. This data is disaggregated by state and holds counters on the number of challenges submitted, withdraw, upheld, and overturned for each provider since the beginning of the BDC challenge process.

For the full technical specification of the availability challenge data, please refer to Sections 4.2 and 4.3 the FCC's Specifications for Data Downloads from the National Broadband Map document [5].

## 2.3 Fabric Challenge Dataset

As described in Section 2.1.1, availability is reported for each known broadband-serviceable location in the nation. This set of known BSLs comprise what is called the fabric layer of the National Broadband Map (NBM). Data on this layer is not publicly accessible via BDC dataset downloads—access is limited through license agreements—although it can be visualized in the map (see Figure 6). Nevertheless, similar to the availability data, the FCC allows for challenges to be submitted to fix inaccuracies in the fabric. These inaccuracies may be due to a location being incorrectly classified as broadband-serviceable or not as well as a location address, building type, or unit count being mistakenly deduced by the FCC. All fabric challenges are evaluated and adjudicated by the FCC. Whenever challenges are upheld, updates are made to the fabric layer. New data on fabric challenges is released monthly, solely in its raw form and disaggregated by state

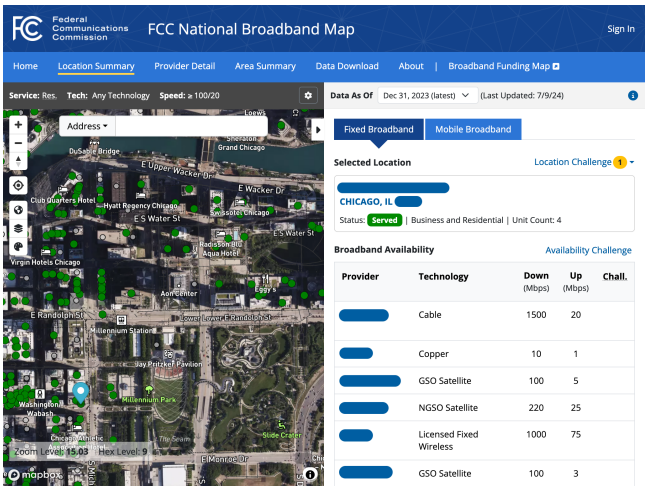


Figure 6: Example of visualizing fabric layer information in the National Broadband Map. The selected location is considered broadband-serviceable and has a building type of mixed use with a total of 4 units.

and adjudication status (either in progress or resolved). Each record in the data contains information on a single challenge. For the challenges in progress, the main fields in these records are a unique identifier for the challenge, the unique identifier for the location, and the reason for the submitting the challenge. For the resolved challenges, outcome-related fields are included. Additional fields may also be present (i.e., non-NULL) indicating corrected metadata regarding addresses, unit counts, building types, geolocation, and whether the location is broadband-serviceable or not.

For the full technical specification of the fabric challenge data, please refer to Section 4.1 of the FCC's Specifications for Data Downloads from the National Broadband Map document [5].

# 3 Challenges and Limitations of the BDC Datasets

In this section, we present the main challenges involved in accessing and analyzing the Broadband Data Collection (BDC) dataset. We contextualize and describe every challenge and, where appropriate, make recommendations to improve data dissemination practices by the FCC as well as bring insights on how other stakeholders may handle these challenges and deal with limitations of the dataset as currently published.

## 3.1 File-Based-Only Data Serving

**Description.** As described in Section 2, the FCC currently serves the BDC dataset purely through file-based downloads. Data is usually disaggregated at the state level and subsequently split into individual files depending on either the access technology or both provider and technology. This data access model is generally neither flexible nor efficient enough. For example, when one may desire to focus on smaller geographical areas (e.g., counties, ZIP code areas, census tracts), this model requires first downloading data on an entire state (typically hundreds of megabytes or several gigabytes in size) and only after that performing filtering actions to select records for the area of interest. As another example, when focusing on data for a provider across states and/or technologies, this access model requires downloading a number of individual files. This number can quickly grow to dozens of files depending on the geographical spread of the provider across states and the variety of technologies it offers. These multistep and multi-file download processes are burdensome to replicate and reproduce, in addition to being fraught with room for human error (e.g., missing files) and wasteful use of computational resources (e.g., data transfer, file storage, and record filtering).

**Recommendation to stakeholders.** Along with the data download portal at the National Broadband Map website, the FCC also offers access to the BDC dataset files via a public API [8]. We recommend using this API to automate file download and make the process easier to replicate and reproduce and less prone to human error.

**Suggestion to the FCC.** We believe that the size of the BDC dataset—hundreds of millions of records amounting to tens of gigabytes of data—calls for making it available not only through file downloads but also via more efficient and flexible data access models. Our suggestion is for the FCC to start serving the BDC dataset via a SQL(-like) query-able API where users could define filters and grouping and inspect overviews of the data before having to download anything, similar to how the FCC Open Data portal offers access to historical Form 477 fillings [6] (see Figure 7). Beyond lowering the barrier for replication and making it easier to detect (and fix) human errors, this would also greatly reduce computational costs [12].

## 3.2 Changes to the Dataset Are Hard to Trace

**Description.** The data contained in the NBM are subject to changes over time. This can be due to many reasons. For example, when providers upgrade their infrastructure, new data may need to be added to reflect the newly covered BSLs. In other cases, as providers migrate to faster and more reliable access technologies, some records may need to be removed for the older technologies. As yet another example, when challenges against data contained in the map are upheld, records may need to be either updated or removed to more accurately

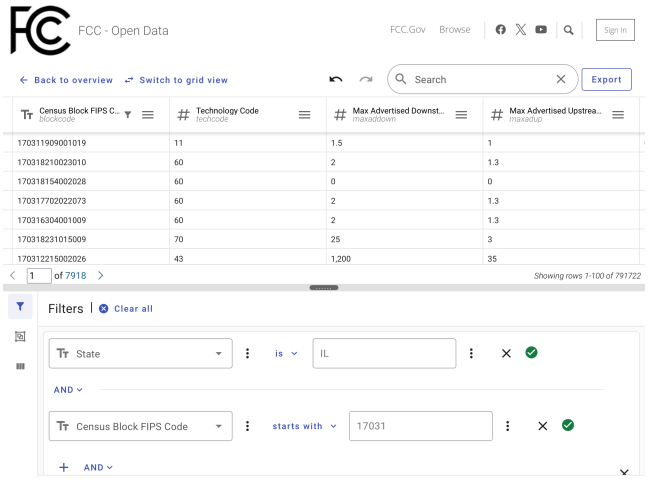


Figure 7: Example of exploring Form 477 data via a SQL-like query-able interface in the FCC Open Data portal.

reflect reality. In view of these issues—and to allow the NBM to capture how Internet access evolves across the nation—the FCC periodically releases new data via what are called "vintages" and continually updates the data contained in those vintages. The issue with this data updating process is that the dataset does not store the date on which records were added (or last modified). Furthermore, availability records that are successfully challenged may be removed retroactively from past vintages. This makes it difficult (if not outright impossible) to answer a range of research questions related to the evolution of Internet access, such as: How fast are providers expanding coverage? What are the areas seeing the most growth? Or the least growth? What specific advertised speeds were past (upheld) challenges disputing?

**Recommendation to stakeholders.** Currently, the best course of action for someone interested in answering questions regarding the evolution of the NBM is to periodically (e.g., weekly) download and store the dataset files whenever they are updated. One intrinsic limitation of this, given the nature of the BDC challenge process, is that past data cannot be retrieved. That is, changes can only be tracked from the first time files are downloaded moving forward in time. We also note that this exacerbates the problems related to resource usage described in §3.1.

**Suggestion to the FCC.** Our suggestion is for three new fields to be added to the availability dataset. The first would be creation_date, the date when the record was first created. The second, is_active, would be a flag indicating whether the availability record is currently active and up-to-date. The third would be inactivation_date, the date when the record became outdated, in case the record is no longer active. Whenever a provider declares new or updated availability, a new record should be created with the creation_date as the current date, the is_active flag set to true, and the inactivation_date set to NULL. If the new record represents an update to existing records, those should be set to inactive and the inactivation_date on them should be set to the current date. This case may arise when a provider updates availability reports to indicate that faster speeds are now offered to customers at the location, for example. Similarly, whenever an availability record is successfully challenged, instead of completely removing it from the dataset, the record should simply be set as inactive and its inactivation_date set. This approach to updating the dataset enables historical data on availability to be queried while at the same time enabling inactive records to be filtered out whenever they are not the focus of analysis.

## 3.3 Summary Data on Availability Is Coarse

**Description.** As described in Section 2, summary data on broadband availability is currently offered at geographical levels ranging from states down to US Census places. Nevertheless, as also described, raw availability data is reported and primarily disseminated on a per-BSL (or per-H3-cell) basis. This means that summaries can in practice be computed for granularity as fine as Census blocks (or even BSLs) as well as for other commonly considered levels such as neighborhoods, ZIP code areas, and Census tracts. However, many of the challenges described in Section 3.1 also apply here since there is the need to download and process multiple possibly large files (covering different states, providers, and technologies)

**Recommendation to stakeholders.** Similarly to the recommendation in Section 3.1, we advise the use of the BDC public API to systematically obtain the necessary files. To deal with the computation requirements (e.g., on memory and storage) that arise especially in this use case, we also recommend loading the data into a database management system (DBMS). DBMSs allow storing data in more compact representations and enable the creation of indexes that simplify data querying and aggregation. Together, these improvements over file-based data storage reduce requirements on storage and memory usage and simplify working within the limitations of these resources.

**Suggestion to the FCC.** The query-able API suggested in Section 3.1 goes some way in addressing these challenges. In that context, users could specify groupings and aggregation functions over the raw data to compute summaries at their desired geographical granularity. At the same time, given the frequency with which studies focus on geographical levels such as those found in the US Census data (i.e., tracts, block groups, and blocks) as well as those related to neighborhoods,

we suggest that the FCC offer summary data at those levels as directly accessible (and query-able) data tables.

### 3.4 No Public Download Access to the Fabric Layer

**Description.** The National Broadband Map (NBM) is composed of two layers: fabric and availability. The availability layer, as described in Section 2, is publicly available for download and compiles Internet service advertisements for each and every (broadband-serviceable) location contained in the fabric layer. The fabric layer contains descriptive metadata related to all locations across the nation. This metadata includes geocoding (e.g., latitude, longitude, Census Block-level FIPS code, postal address), building type (e.g., residential, enterprise), number of units, intended land use, and—most crucially—whether the location is considered broadband-serviceable or not. Contrasting with the availability data, the fabric data is not publicly accessible but subject to license agreements. This situation hinders analysis efforts that require taking sample or population characteristics and sizes into account, which can be exemplified by questions such as: What is the number of residential units lacking (served-level) broadband access in an area? How about enterprise units or community anchor institutions (CAIs)? Is lack of access concentrated around rural or urban areas? Does proximity to CAIs reflect easier access or better performance?

**Recommendation to stakeholders.** Although fabric data is not publicly available to its full extent, some summary counters on this data either (a) are available along with the availability summaries (e.g., number of BSLs in an area) or (b) can be estimated from raw availability data (e.g., counting the number of unique location IDs that fall within each area). The downside of the summary data is that it is general (e.g., no distinction between different BSL types) and limited to large geographical areas (e.g., states, congressional districts, Census places; See §3.3). The main limitation with estimating counters from raw availability data is that it can only consider locations with at least one Internet access offering. This limitation is especially problematic in "broadband desert" areas, where availability is minimal or nonexistent, since those areas are virtually invisible through the availability perspective and, generally, the most in need of policy intervention. Hence, our general recommendation is to rely on summary data whenever possible. We note that other sources of data (e.g., US Census Demographics, USDA Rural-Urban Continuum Codes) may be applied to cope with the lack of access to the fabric layer, though these represent a layer of indirection between the availability and BSLs as determined by the FCC that should be carefully considered.

**Suggestion to the FCC.** Understanding that some metadata contained in the fabric layer is legally owned by a third party (and not by the FCC), we suggest two approaches for enabling the general need for the data. First, the FCC could publicly release raw data that omits sensitive metadata fields and only includes those locations considered to be broadband-serviceable. Second, similar to what is done for the availability layer, the FCC could offer data summaries on the fabric layer, while also including finer granularity levels (e.g. Census blocks and H3 resolution 9). For general use, the crucial metadata fields to include would be the location ID, the location Census block FIPS code, the location ZIP code, the building type, and the unit count.

### 3.5 Lack of Data Before June 2022

**Description.** Historically, the FCC has collected and disseminated broadband availability as reported by Internet service providers using Form 477. This form had to be filled out twice a year by providers and shared with the FCC. Coverage in the Form 477 was reported per Census block, with providers sharing the number of units that could subscribe to its services and the maximum speed offered to any one unit in the block. In March 2020, under the Broadband DATA Act, the US Congress directed the FCC to build a more accurate map of broadband deployment and technology availability across the nation. This led the FCC to collect availability data through the Broadband Data Collection (BDC) program and disseminate it through the National Broadband Map (NBM) portal, with the first release reflecting deployment as of the end of June 2022. Among others, the main improvement of the BDC dataset is that data is more fine-grained as it must be reported per location considering all BSLs in the fabric layer (i.e., the common dataset of serviceable locations defined by the FCC). However, aggregated at the Census-block level, the BDC dataset could be contrasted with past datasets based on Form 447 reports. Nevertheless, the two datasets are made publicly available on separate websites and through distinct interfaces and APIs, which makes any comparison more difficult, requiring significant pre-processsing.

**Recommendation to stakeholders.** Considering that in both cases data is made publicly available, the general recommendation is to rely on automated and reproducible approaches to integrate these datasets.

**Suggestion to the FCC.** Given the importance of understanding the evolution of broadband deployment across the nation, we suggest that the FCC lower the barriers for comparison between these important datasets. There are two complementary approaches towards this goal. The first would be integrating the download of Form 477 data into the BDC web-

site and API (with the appropriate disclaimers on the change in data granularity). The second is to compute and publish availability summaries for BDC data at the granularity offered by Form 447 data (i.e., Census block granularity). This would eliminate the need for repeated efforts to compute these summaries by other stakeholders, as discussed in Section 3.3.

# 4 Conclusion

In this paper we shared our experience and lessons learned from exploring and analyzing the Broadband Data Collection (BDC) program datasets. We identified several challenges and limitations that one may face when trying to analyze these datasets. Amongst those are file-based-only data serving, loss of past data related to challenges and updates, and lack of public download access to the fabric layer. To help other researchers, practitioners, and policymakers, where possible, we recommended ways to overcome or handle these challenges and limitations. Furthermore, we also made suggestions on ways that the FCC can change its data publication, serving, update, and summarization processes to facilitate and encourage engagement by all interested parties. These suggestions included, for example, serving the datasets via a query-able API, addition of metadata fields to track data updates, and a selective model for publishing fabric layer data.

**Acknowledgments.** This research was funded by NSF awards CNS-2223610, CNS-2213821, CNS-2319603, and CNS-2224687. The work was carried out as a project of the University of Chicago's Data Science Institute data clinic, and we would like to thank our data clinic co-mentor Tim Hannifan and the students who worked on this project (ordered by last name): Damian Dhillon, Aaron Haefner, Angelie Miranda, Ridhi Purohit, Neha Sadasivan, Elena Smyslovskikh, Shwetha Srinivasan, and Ruoyi Wu. Their valuable efforts in exploring the BDC datasets have allowed us to develop insights on how to make working with this data easier and more effective.

# References

[1] Environmental Systems Research Institute. *ESRI Shapefile Technical Description*. 1998. URL: https://support.esri.com/en-us/technical-paper/esri-shapefile-technical-description-279 (visited on 07/23/2024).

[2] Federal Communications Commission (FCC). *Rosenworcel Establishes Broadband Data Task Force*. 2021. URL: https://www.fcc.gov/document/rosenworcel-establishes-broadband-data-task-force (visited on 07/23/2024).

[3] Federal Communications Commission (FCC). *FCC National Broadband Map*. FCC National Broadband Map. 2022. URL: https://broadbandmap.fcc.gov (visited on 02/21/2024).

[4] Federal Communications Commission (FCC). *FCC Releases New National Broadband Maps*. 2022. URL: https://broadbandmap.fcc.gov (visited on 07/23/2024).

[5] Federal Communications Commission (FCC). *Broadband Data Collection: Specifications for Data Downloads from the National Broadband Map*. 2023. URL: https://us-fcc.app.box.com/v/bdc-data-downloads-output (visited on 02/21/2024).

[6] Federal Communications Commission (FCC). *FCC Open Data - Fixed Broadband Deployment Data: June 2021 Status V1*. 2023. URL: https://opendata.fcc.gov/Wireline/Fixed-Broadband-Deployment-Data-June-2021-Status-V/jdr4-3q4p/about_data (visited on 07/23/2024).

[7] Federal Communications Commission (FCC). *About the Fabric: What a Broadband Serviceable Location (BSL) Is and Is Not*. 2024. URL: https://help.bdc.fcc.gov/hc/en-us/articles/16842264428059-About-the-Fabric-What-a-Broadband-Serviceable-Location-BSL-Is-and-Is-Not (visited on 07/23/2024).

[8] Federal Communications Commission (FCC). *Broadband Data Collection: National Broadband Map Public Data API Specifications and Instructions*. 2024. URL: https://us-fcc.app.box.com/v/bdc-public-data-api-spec (visited on 07/23/2024).

[9] Internet for All. *BEAD Challenge Process Policy*. 2023. URL: https://internet4all.gov/bead-challenge-process-policy (visited on 02/21/2024).

[10] National Telecommunications and Information Administration (NTIA). *The Biden-Harris Administration Launches $45 Billion "Internet for All" Initiative to Bring Affordable, Reliable High-Speed Internet to Everyone in America | National Telecommunications and Information Administration*. 2022. URL: https://www.ntia.gov/press-release/2022/biden-harris-administration-launches-45-billion-internet-all-initiative-bring (visited on 02/21/2024).

[11] Open Geospatial Consortium. *GeoPackage: An Open Format for Geospatial Information*. 2024. URL: https://www.geopackage.org/ (visited on 07/23/2024).

[12] Mark Raasveldt and Hannes Mühleisen. "DuckDB: an Embeddable Analytical Database". In: *Proceedings of the 2019 International Conference on Management of Data*. SIGMOD '19. Amsterdam, Netherlands: Association for Computing Machinery, 2019, pp. 1981–1984. ISBN: 9781450356435. DOI: 10.1145/3299869.3320212. URL: https://doi.org/10.1145/3299869.3320212.

[13] Yakov Shafranovich. *Common Format and MIME Type for Comma-Separated Values (CSV) Files*. RFC 4180. Oct. 2005. DOI: 10.17487/RFC4180. URL: https://www.rfc-editor.org/info/rfc4180.

[14] Uber Technologies. *H3 - Geospatial Indexing System*. 2024. URL: https://h3geo.org/ (visited on 07/23/2024).

[15] United States 116th Congress (2019-2020). *S.1822 - Broadband DATA Act*. 2020. URL: https://www.congress.gov/bill/116th-congress/senate-bill/1822/text (visited on 07/23/2024).

[16] United States 117th Congress (2021-2022). *H.R.3684 - Infrastructure Investment and Jobs Act*. 2021. URL: https://www.congress.gov/bill/117th-congress/house-bill/3684 (visited on 07/23/2024).