

**Keith Wardrip<sup>1</sup>**

**Abstract**

This article details the data-specific challenges encountered by the National Low Income Housing Coalition in its efforts to create a preservation catalog, defined here as a comprehensive database of all federally assisted housing units with details on the location, condition, and the federal subsidy or subsidies that make rents affordable. Relying primarily on experiences with the publicly available HUD datasets, specific examples illustrate problems with database integration and questionable data quality. More importantly, this article includes suggested workarounds for these obstacles. A clear policy recommendation that emerges from this research is that HUD, USDA and other appropriate agencies work to fully implement a system of unique IDs for all projects in the federally subsidized stock.

**A National Preservation Catalog**

Today, HUD's Picture of Subsidized Households (Picture), which includes information on public and multifamily housing and Low Income Housing Tax Credit (LIHTC) projects, is the closest approximation to a national preservation catalog that is currently available. Picture does not cover the entire income-restricted housing stock, however, because it excludes assisted rural housing and HOME-funded developments. Even if its coverage of the affordable housing programs were complete, the current version, released in late 2006, reflects a snapshot of the subsidized housing stock taken in the year 2000. Thus, for a variety of reasons, Picture is less than ideal as a preservation catalog

Separate from Picture, HUD does provide more current databases for projects made affordable through the project-based Section 8 program, FHA insured or subsidized mortgages, housing for the elderly and disabled (Sections 202 and 811, respectively), and LIHTC developments. Depending on the program, these datasets are updated anywhere from monthly to annually<sup>2</sup> and are publicly available on either the HUD or HUDUSER website. Data on projects made affordable through HUD's HOME program and USDA's Rural Housing Service (e.g., Section 515) are not readily available to the general public in a database format, but NLIHC has successfully acquired these data in the past.

As part of an investigation into what it would take to create a national preservation catalog, NLIHC attempted to compile these various datasets into a single database. This article describes specific obstacles that were encountered during the development of preservation catalogs for Florida and the District of Columbia, as well as proposed workarounds for these challenges. Closing comments focus on actions that should be taken by HUD and other data providers to largely resolve the problems described here.

---

<sup>1</sup> The author would like to thank the John D. and Catherine T. MacArthur Foundation for supporting this research and Danilo Pelletiere and Rebecca Warden for their comments on previous versions of this article.

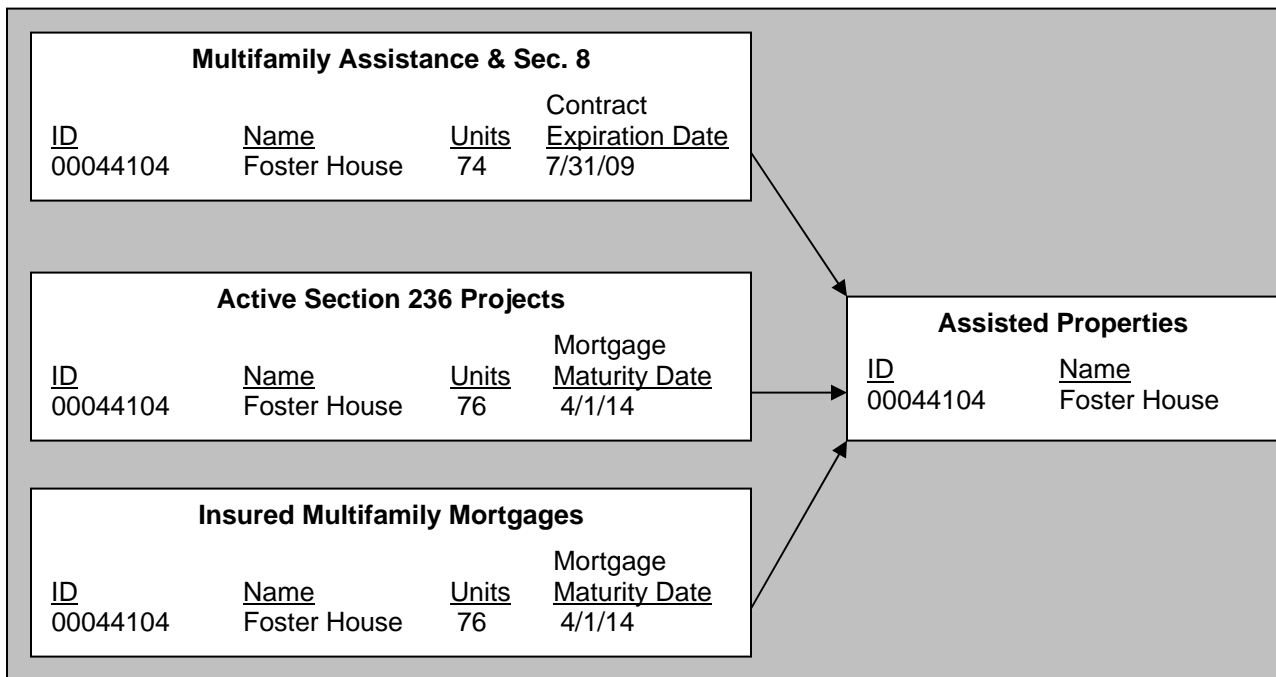
<sup>2</sup> Despite annual updates, there is a three-year lag between when a project is placed in service and when it appears in the LIHTC dataset (i.e., the 2006 version included only projects placed in service through 2003).

### Preservation Catalog Design

A preservation catalog is predicated on the assumption that the ideal database structure involves the creation of a table of assisted properties – each represented by only one record – that can be linked to or associated with data for the federal subsidy or subsidies that make each project affordable.

The tables provided by HUD for project-based Section 8 contracts, insured multifamily mortgages, Section 236, and Section 202/811 properties contain a unique ID field that in the absence of error allow the same project to be identified in each table. This unique ID facilitates the creation of an assisted properties table with one record for each project, as Figure 1 illustrates for Foster House in D.C.

**Figure 1: Creating a Single Table of All Assisted Properties**



With a unique ID to make these linkages possible, all of the pertinent information related to each subsidy can then be displayed in an easy-to-read report (Figure 2).

**Figure 2: Report Linking the Property Table with Subsidy Data**

Preservation Catalog		
<i>Foster House</i>		
Assisted <u>Units</u>	End of <u>Affordability</u>	<u>Data Source</u>
74	7/31/09	Multifamily Assistance & Sec. 8 Contracts
76	4/1/14	Active Section 236 Projects
76	4/1/14	Insured Multifamily Mortgages

**Data Issues and Challenges**

Unfortunately, this project identifier is not present in all relevant databases. Even when it is present, however, it is not always sufficiently accurate to use without careful scrutiny of the data. This section illustrates specific problems of data structure and quality encountered in the datasets.

Lacking Unique IDs

Figure 3 brings together records for the same project from three databases and perfectly illustrates one of the obstacles rooted in the lack of a consistent project identifier.

**Figure 3: No Consistent ID Across Datasets**

Property ID/ HUD ID	Name	Address	Database
800024865	Woodbury Apartments	2234 SAVANNAH TERR SE	Multifamily Assistance & Section 8 Contracts
DCB2001005	SAVANNAH RIDGE/WOODBURY VILLAGE	2224 SAVANNAH TER SE	Low Income Housing Tax Credit Projects
	SAVANNAH RIDGE APARTMENTS	2202-2245 SAVANNAH TERR & 3225-3335 22ND	HOME Projects (provided by HUD)

This project has a different ID in the Multifamily Assistance & Section 8 Contracts database than it does in the LIHTC database, and in the HOME dataset it has no unique ID at all. Though admittedly not a perfect match, data provided in the name and address fields indicate that these three records correspond to the same housing development. Telephone calls, internet research, and a windshield survey were necessary to confirm this suspicion when the presence of a consistent ID across the datasets would have streamlined the process and removed all doubt.

Data Quality Issues Across Datasets

As essential as they are to facilitating the process of data integration, unique ID fields as they are currently populated in HUD datasets are not failsafe. Of the roughly 170 projects in D.C. and Florida that are cross-subsidized by project-based Section 8 and FHA mortgage insurance, 12 were found to have incorrect or missing IDs that prevented the initial synthesis of information from the two sources. These mistakes and omissions, which may be the result of project refinancing and a lack of follow-through in database maintenance, were discovered only after carefully scrutinizing the data record by record and were also observed in the Section 236 and Sections 202/811 datasets.

When project IDs function as they are intended and allow users to identify the same project in multiple datasets, they often highlight inconsistencies in the data. Of the 44 projects in D.C. with both a project-based Section 8 contract and an insured mortgage, the datasets commonly listed different project addresses. Some were close approximations of the same location (e.g., 6<sup>th</sup> & H Sts., NW versus 800 6<sup>th</sup> St NW) but others lacked street numbers or disagreed on street names (e.g., Corcoran Street, NE versus 1050 Mount Olivet Rd, NE). Conflicting addresses are particularly problematic if the user intends to evaluate a jurisdiction’s assisted housing stock within a broader neighborhood context.

Focusing on a variable much easier to quantify, the two datasets disagreed on the total number of units in the same development in 17 of the 44 cases in D.C., with the difference ranging from one to 407 units. In Florida, the total number of units differed in 14 of 128 cases.

Data Quality Issues Within Datasets

These case studies indicate that it is not unusual for a single publicly available dataset to contain more than one record – each with different data – for the same project. The issue of multiple project records is a serious one because in order to ensure that the resulting database has accurate information, additional, often time-intensive, research is required to validate one record and delete the other. This is a particularly important process when data related to a project’s period of affordability are called into question by multiple and contradictory records.

Multiple records discovered in the Insured Multifamily Mortgages and the Section 236 databases often appear to involve a project refinancing without the subsequent removal of the previous, outdated record. As Figure 4 illustrates, redundant records in the LIHTC dataset often have conflicting ‘year placed in service’ dates.

**Figure 4: Duplicate Records in the LIHTC Dataset**

HUD ID	Name	Address	Units	Year Placed in Service
FLA0000077	LAKEWOOD TERRACE	1315 W 14TH ST	132	1989
FLA1994024	LAKEWOOD TERRACE APARTMENTS	1315 W 14TH ST	132	1994

Correspondence with Abt Associates, Inc., the contractor for the LIHTC database, indicates that multiple

listings are sometimes appropriate because they reflect different project phases with separate tax credit allocations. Alternatively, these records could indicate that a single project received multiple allocations of tax credits, or they may simply be the product of imperfect database oversight. This ambiguity complicates the integration of the LIHTC file into any preservation catalog.

The HOME dataset procured from HUD’s Office of Community Planning and Development was particularly prone to duplicate records, which is likely due to data collection procedures. Local HOME offices are responsible for entering and updating all projects receiving HOME funds, and the varying levels of technical expertise may explain the prevalence of duplicate entries.

## Workarounds

The Coalition developed several effective, albeit imperfect, methodologies for tackling the challenges described above. The most important issue to resolve, as well as the easiest to tackle methodologically, is the synthesis of the datasets lacking unique IDs (e.g., LIHTC, HOME, RHS) into a table of assisted properties. Figure 5 lists the records for a project with three such subsidies. For heuristic reasons, the example is a hypothetical project based on a composite of characteristics and problems found within the D.C. data.

**Figure 5: Consolidating Datasets Lacking Unique IDs**

Assisted Properties				
<u>ID</u>	<u>Name</u>	<u>Total Units</u>	<u>Address</u>	<u>Zip Code</u>
0017	Main Street Apartments	553	123 Main St	20005
	Main St. Apts.	337	123 Main Street	20005
DC46A	Apartments on Main	553	Main St. & 2nd St	20005

Because there is no one piece of information that automatically identifies these three records as the same property, it must be assumed in the initial stages of creating the assisted properties table that they are different projects. The result in this case, however, is the triplicate listing of Main Street Apartments in the catalog. The following three strategies are suggested by the Coalition to identify duplicate cases like these in the table of assisted properties and consolidate them into one record.

1. Sorting. The first and most effective strategy for identifying duplicate records is simply to sort the table by project name. Doing so would call attention to the redundancy of the first two records, above, and allow the database administrator to merge them into one. This process should be repeated using the address field because, as indicated in several earlier examples, properties can change names without it being reflected in the data, while addresses are assumed to be more stable over time.
2. Querying. After sorting, it is helpful to query the table for records that have similar geographic identifiers and project sizes. For example, one could query the data for all projects in the same city or zip code reporting an identical number of units. This can reveal duplicate records with slightly different names or addresses that may be overlooked by simply sorting the data (e.g., the first and third records in Figure 5).
3. Geocoding. Because the size of some large assisted projects – ranging from several buildings to distinct “neighborhoods” – can render the address field less reliable, it can be helpful to geocode the records in a GIS and closely scrutinize projects in close proximity to one another. Records of projects with slightly different names and addresses can escape the aforementioned workarounds and still be identified spatially through geocoding.

Internet research, phoning property owners and managers, and pounding the pavement may be necessary to resolve data conflicts that persist after the techniques above have been employed, but these three processes go a long way toward overcoming the lack of a

unique project ID. Where this project-specific research is necessary, experience has shown that conflicting information should not be consolidated or deleted until the discrepancies are resolved with a high level of certainty.

To improve the accuracy and facilitate the maintenance of the preservation catalog, the Coalition is pursuing the notion of establishing a network of “project monitors.” As currently envisioned, each project in a preservation catalog would be associated with a local organization (e.g., community development corporation, local affordable housing office, faith-based group, tenant association) with firsthand knowledge of that project. An open exchange of information between the catalog’s host and the project monitors would allow the latter to not only verify data provided by HUD but also update it regularly as circumstances change.

### **Critical but Unavailable Data**

Thus far, this article has focused on the challenges and workarounds associated with data that HUD currently makes available to the public. Any discussion of efforts to create a preservation catalog would be incomplete, however, if it ignored additional data that are essential but currently unavailable.

First and foremost, a preservation catalog needs to include an indication of which properties are at-risk of losing their subsidy because the owner has decided to opt-out of a Section 8 contract or prepay a subsidized mortgage. Owners must notify their local HUD office and their tenants at least 12 months before prematurely ending the period of affordability, but this information is not collected – or at least not disseminated – nationally. As an illustration of what is possible, USDA operates a searchable online database for nonprofits that identifies rural housing projects for which mortgage prepayment is expected, but no parallel system exists for projects subsidized by HUD.

Also highly sought after by housing preservation groups, HUD’s Real Estate Assessment Center (REAC) scores reflect the physical and financial soundness of subsidized properties and can provide an early warning that a housing development’s affordability status is in jeopardy. As important as they are to those interested in preserving affordable housing, however, HUD has not made these scores available to the public in any kind of comprehensive national database.

### **Recommendations to Data Providers**

As the agency primarily responsible for the construction and operation of assisted housing in the U.S., HUD should also be responsible for providing a comprehensive, regularly updated, and high-quality preservation catalog that is national in scope. This catalog should highlight projects known to be at risk of losing their federal subsidy through an owner’s decision to either opt-out of a Section 8 contract or prepay an FHA mortgage. Such a product could stand alone in areas where the capacity for housing advocacy is limited, or it could serve as the foundation for an agency wishing to integrate locally funded projects and local knowledge into a national infrastructure.

Barring this development, HUD should coordinate with USDA and state housing finance agencies in an effort to assign all assisted housing developments a unique project identifier. Such a seemingly simple piece of information would make the task of assembling a preservation catalog from a vast array of disparate databases a much more efficient and effective process, and it would render moot many of the workarounds identified in this article. Additionally, HUD should engage the community of housing advocates to determine what other data it might provide at little cost to strengthen affordable housing preservation efforts (e.g., opt-out and prepayment notices, REAC scores, etc.).

### **Conclusions**

While there is clearly room for improvement in the provision of data on the assisted housing stock, the purpose of this article is to encourage researchers and local organizations to fully utilize the datasets that do exist. It is hoped that by describing the data challenges and the suggested workarounds, this discussion will help other database developers and end-users maximize the utility of the available data and ultimately use the information to preserve the nation's affordable housing stock.