yext

The Definitive Guide to Duplicate Listings

By **Andrew Shotland**, Founder of **LocalSEOGuide**

Table of Contents

Introduction

In the Local SEO business, we spend a lot of time dealing with duplicate business listings. Duplicate records of your business appearing throughout the Local Search ecosystem are bad news for a variety of reasons, not the least of which is they could be costing your business a lot of time and money. As an industry, it behooves us to collectively solve this issue.

Why Are Duplicate Listings Bad?

Here's a list of some of the bad things that can happen when you have duplicate listings:

1. Customer Confusion & Frustration

When someone searching for a local service encounters a duplicate listing with incorrect data or multiple listings for the same business, it can cause confusion and frustration. In 2013 we estimated that incorrect business listings data cost businesses $10 billion per year.

2. Customer Communication Challenges

Many local search sites allow business owners to respond to customers via their claimed listings. In cases where there are multiple listings for a location, it can be a challenge for the business to find all of them to claim and manage the customer interaction across multiple versions of the same listing. And if a listing is being used to share content like blog posts, etc., the business is then required to post via each dupe or else focus on only one and hope that their customers find the listing with the updated content.

3. Google Rankings Issues

Google uses business listing information from a variety of sources such as online yellow pages publishers and data aggregators as part of its local rankings algorithm. If you have duplicate listings or if your listings data is not consistent across the various data sources, you may have problems ranking well for relevant local search queries.

4. Social Sharing Spread Thin

Many local search sites allow customers to add social data to business listings (e.g. check-ins, photo & video uploads, reviews, etc.). Local search engines often use the presence of this kind of content in ranking algorithms for their internal search engines and Google definitely likes business profiles that are regularly generating content. Business profiles with a lot of this kind of content also tend to get shared more often than those with less. If there are multiple listings for a business, this social activity might get spread out amongst several dupe profiles, making it harder for any one profile to achieve a critical mass of content.

Common Types of Duplicate Listings

So dupes are bad. But what causes them? Let's examine the common types and how they are created:

1. Self-Created Dupes

This happens over time when a business does not have a cohesive strategy to deal with their business listing. Typical self-created dupes happen when different parties in a business claim or add profiles to various directories and data suppliers without knowing that someone else in the organization has done this already. This also can happen if you are using a third-party tool to add a listing to various data aggregators and publishers and the tool does not effectively detect downstream duplicates.

2. Aggregator-Created Dupes

The business listing data aggregators gather information from a variety of sources (sometimes thousands) to determine a business' name, address, phone number, etc. The problem is that these aggregators turn out to be not that great at matching up the records from various sources (it's a tough job which even Google struggles with) and during the matching process, more duplicates can be created which in turn can have an ongoing pollutant effect downstream at the the publisher level.

3. Publisher-Created Dupes

At the publisher level itself, duplicates run wild as publishers have lax matching and data cleanup policies. Part of the problem is that local directory publishers have business models that are at odds with the average business trying to clean up their dupes, particularly if they are doing it for SEO. The publisher typically is looking to get paid for improving the presence of a local business on their network and is not as concerned with how the business appears in Google. In fact, I would argue a directory publisher has an incentive for the business to not show up well in Google, because then it will need to buy more leads from the publishers. So getting the publishers to police and clean up their own bad data can be slow and ineffective.

4. Cross-Pollination Dupes

Because of crawling practices, there is a constant collision problem in the ecosystem where a publisher or aggregator crawls their way back into more dupes. For example, an aggregator sends a two dupe listings to a publisher, even if the aggregator fixes the dupe issue, if it relies on web-crawling as a source and it crawls the publisher's site, the dupes could end up coming back to haunt you. Kind of like zombies.

How to Solve the Problem

Let's take a deeper look at how dupes are technically created at the Aggregator and Publisher levels and how the problem can be solved.

Every local search publisher works with dozens of local data sources. Publishers put the data from these sources through two steps—**Cluster** and **Conflate**—to arrive at a set of data (the view) that appears to an end user.

Let's take fictional publisher "Bingo". Say they pull in data from 8 sources (usually it is ~50). Bingo runs its data compilation process for a fictional business: "Joe's Pizza", in Fargo ND. First, Bingo runs "Cluster".

The Cluster

The objective of the cluster process is to identify which records in each source apply to a particular location. This is easy with the human eye. (though painstaking as you'd have to comb through billions of records). But say someone does just that for Joe's Pizza, they find:

SOURCE	NAME	ADDRESS	PHONE	WEBSITE	OTHER
InfoGroup (ID: 34131)	Joe's Pizza - Fargo	585 31st Street, Fargo ND	(703) 456-1234	-	-
InfoGroup (ID: 2141)	Joe's Pizza & Pasta	-	(888) 321-4991	-	-
Localeze (ID: ABC913)	Joe's Pizza	585 31st Street, Suite 47, Fargo ND	(703) 456-1234	joespizza.com	-
Axciom (ID: 913119)	Joe's Pizzeria	-	(703) 456-1234	joespizza.com	-
IRS (ID: 99-13-019)	Joe's Pizza, LLC	585 31st St., Fargo ND	(701) 555-5055	-	-
BBB (ID: WISE4910)	Joe's Pizza of Fargo Inc.	585 31st Street, Suite 47, Fargo ND	(703) 456-1234	joespizza.com	-
Web Crawling (ID: 5011011)	Joe's Pizza	585 31st Street, Suite 47, Fargo ND	(703) 456-1234	joespizza.com	-
Business Claim Data (ID: 4990)	Joe's Pizza	585 31st Street, Fargo, ND	(888) 321-4991	-	-
User Generated Data	Joe's Pizza	585 31st Street, Suite 47, Fargo ND	(703) 456-1234	-	Marked as Closed

It's much harder for a computer to do this. Every source has slightly different information about Joe's Pizza. The human eye easily recognizes that the IRS record and the first InfoGroup record are probably the same, despite different names and addresses. But publishers must build algorithms that analyze which records are the same across and cluster them.

Bingo runs their cluster process for Joe's Pizza. Let's say they are perfect and the output is the same as the human eye. The next step is Conflation.

Conflation

Now, Bingo must conflate the data from each source to decide which name, address, phone, etc. to show in the view. They do this by ranking each source at the element level. The data that has the highest rank wins.

Bingo's rank by element for each source:

SOURCE	NAME	ADDRESS	PHONE	WEBSITE
InfoGroup	2	8	1	2
Localeze	1	5	8	3
Axciom	4	4	6	7
IRS	7	1	5	6
BBB	5	3	2	5
Web Crawling	6	8	3	1
Claim Data	8	2	4	8
User Generated Data (MapMarker)	3	7	7	4

This output is the following:

Joe's Pizza *(Localeze, ABC913)*
31st St., Fargo ND *(IRS, 99-13-019)*
(703) 456-1234 *(InfoGroup, 34131)*
joespizza.com *(Web Crawling, 5011011)*

This process worked perfectly because Bingo's cluster worked flawlessly—it caught a dupe (in the IG data) and more importantly, was able to match up a bunch of records with slightly different information.

Unfortunately, things often don't work out so perfectly.

Duplicates Come From Either Bad Source Data Or Publisher Cluster Mistakes

Let's go back to our example about Joe's Pizza. Once again, here's what the human eye picked up from each source. Importantly, note that the human picked up two InfoUSA records (34131 and 2141) as the same business and knew to *merge* them as *one single record*.

It's hard to train a computer to do the same thing! The records have different names and phone numbers. There is nothing to match them up except intuition. Say Bingo's match algo is not perfect and does not cluster InfoGroup 2141 with the others. Annoyingly, this creates a duplicate listing for Joe's Pizza, which appears on Bingo.

So now there are two listings for Joe's Pizza:

Correct Listing: *(Bingo ID: 1910991)*

Joe's Pizza, Fargo *(Localeze, ABC913)*
585 31st Street Suite 47, Fargo ND *(IRS, 99-13-019)*
(703) 456-1234 *(InfoGroup, 34131)*
joespizzafargo.com *(Web Crawling, 5011011)*

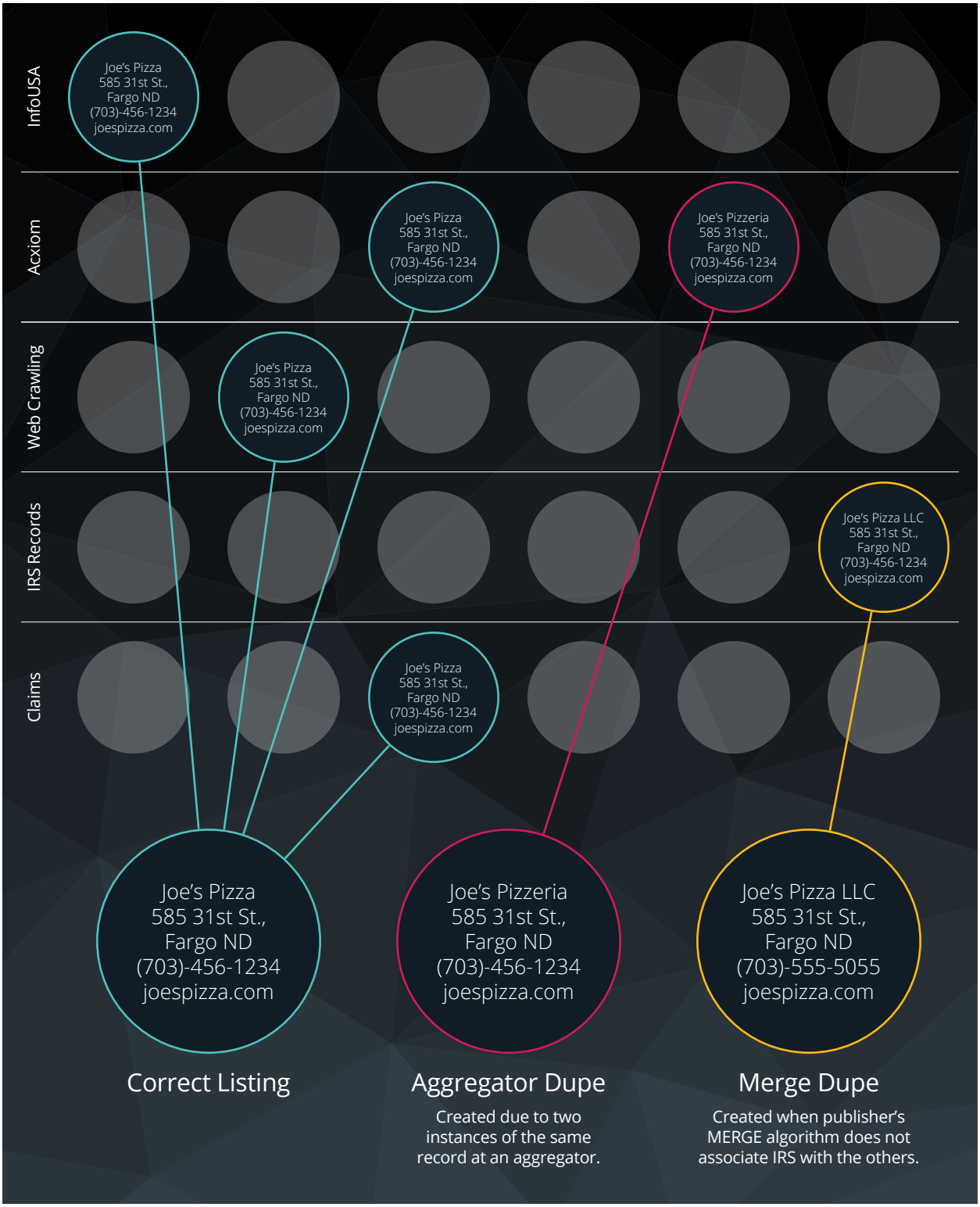
Duplicate Listing: *(Bingo ID: 6010987)*

Joe's Pizza & Pasta *(InfoGroup, 2141)*
585 31st Street, Fargo ND *(Business Claim Data)*
(888) 321-4991 *(InfoGroup, 2141)*

As discussed, duplicate listing can have implications for a business' online presence, and so Joe's Pizza sets off to eliminate it.

SOURCE	NAME	ADDRESS	PHONE	WEBSITE	OTHER
InfoGroup (ID: 34131)	Joe's Pizza - Fargo	585 31st Street, Fargo ND	(703) 456-1234	-	-
InfoGroup (ID: 2141)	Joe's Pizza & Pasta	-	(888) 321-4991	-	-
Localeze (ID: ABC913)	Joe's Pizza	585 31st Street, Suite 47, Fargo ND	(703) 456-1234	joespizza.com	-
Axciom (ID: 913119)	Joe's Pizzeria	-	(703) 456-1234	joespizza.com	-
IRS (ID: 99-13-019)	Joe's Pizza, LLC	585 31st St., Fargo ND	(701) 555-5055	-	-
BBB (ID: WISE4910)	Joe's Pizza of Fargo Inc.	585 31st Street, Suite 47, Fargo ND	(703) 456-1234	joespizza.com	-
Web Crawling (ID: 5011011)	Joe's Pizza	585 31st Street, Suite 47, Fargo ND	(703) 456-1234	joespizza.com	-
Business Claim Data (ID: 4990)	Joe's Pizza	585 31st Street, Fargo, ND	(888) 321-4991	-	-
User Generated Data	Joe's Pizza	585 31st Street, Suite 47, Fargo ND	(703) 456-1234	-	Marked as Closed

Figure: Duplicate Listings are Created by Aggregators and Publishers



Getting Rid of the Data at the Source is not Enough

This example above illustrates a duplicate problem that can be corrected at the source. If the source (InfoGroup in this case) clusters the records, in theory, the problem should be solved.

This is extremely challenging and doesn't really solve the problem.

It's challenging because you first need to guess all of the sources ingested by a publisher. The problem is publishers generally don't name all the sources they use. It's their black box. You must also guess the culprit source (or sources) perfectly. If there's just one mistake, the next time the publisher runs their match process, the dupe will be created and start appearing again. Or if a dupe makes it into the wild, and the publisher crawls it, you'll be unable to guess the source. This happens all the time.

Worse, its impossible to know what truly causes a dupe. It's not always bad source data. It's often the publishers' cluster process.

Look at the record for Joe's supplied by Axciom and their own Web crawling files:

SOURCE	NAME	ADDRESS	PHONE	WEBSITE
Axciom (ID: 913119)	Joe's Pizza	-	(703) 456-1234	joespizzafargo.com
Web Crawling (ID: 5011011)	Joe's Pizza & Pasta	585 31st Street, Suite 47, Fargo, ND	(703) 456-1234	joespizzafargo.com

The data is close, but the name and addresses are different. Bingo runs their cluster process. Say their algo determines that each record is a distinct entity, so a duplicate is created. Two listings appear on Bingo.

In this case, going to Axciom won't eliminate the duplicate. Axciom doesn't have two records. And they have the right data. You'd have to know how to correct the web crawling source, which you don't have access to. An imperfection in Bingo's merge algorithm causes the dupe.

Dupes can be born out of either bad source data (duplication or bad data within a source) or an imperfect cluster.

Dupes Must be Solved at the Publisher Level

The key to solving the problem is addressing duplicates at the publisher level. That way you nip them right before they surface.

An Overlay with a "Hide" or "Redirect" Flag is the Solution to Duplicate Listings.

This is possible at publishers who have implemented an overlay with a "hide flag". With a hide flag overlay, when the merge process is run, a publisher knows to "hide" the duplicate record from their view in question:

> **HIDE** *(Bingo ID: 6010987)*
> Joe's Pizza & Pasta *(InfoGroup, 2141)*
> 585 31st Street, Fargo ND *(Business Claim Data)*
> (888) 321-4991 *(InfoGroup, 2141)*

An overlay with HIDE trumps this record from appearing to the end user, solving the problem.

This happens differently at publishers with stable or dynamic listings. Every local search publisher has either a stable listing ID (meaning they keep a constant identifier for all listings they've built) or unstable listing ID (meaning they regenerate a listing ID each time they rebuild their view).

Duplicate Suppression at Publishers with a Stable Listing ID

It's easier to handle a duplicate for a publisher with a stable listing ID. Assuming they are willing, you can tell them the ID of the listing that's a dupe.

Say Bingo has a stable listing ID. If they have implemented the HIDE flag overlay architecture, just tell them their own ID of which listing to "hide" and this happens the next time they run their merge. (Tip: The best architecture is actually a redirect. This way, all the traffic from a dupe ends up going to the originally intended listing)

Duplicate Suppression at Publishers with a Dynamic Listing ID

This is a lot trickier because the ID of the duplicate changes every time a publisher runs their merge. In this case, you need to provide a way for the publisher to identify the duplicates created by their merge so they know to normalize the dupe data in a way that's picked up by their merge.

Back to Joe's Pizza's duplicate listings:

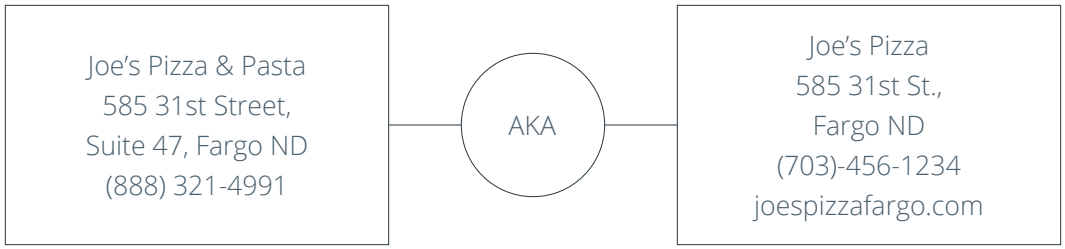
Correct Listing: *(Bingo ID: 1910991)*

Joe's Pizza *(Localeze, ABC913)*
585 31st St., Fargo ND *(IRS, 99-13-019)*
(703) 456-1234 *(InfoGroup, 34131)*
joespizzafargo.com *(Web Crawling, 5011011)*

Duplicate Listing: *(Bingo ID: 6010987)*

Joe's Pizza & Pasta *(InfoGroup, 2141)*
585 31st St., Ste 4, Fargo ND *(Business Claim Data)*
(888) 321-4991 *(InfoGroup, 2141)*

To correct a dupe, you need to correct the problem in the merge. So you provide them the dupe data in an "AKA" format, so they can apply this to their merge algorithm. Like this:



The next time Bingo runs their merge, and encounters the data on the left, they know to merge it into one with the data on the right.

An Overlay Solves the Problem

The elegance of the "Overlay Hide" architecture is that Bingo doesn't reveal their sources. They gain intelligence on what data's good and what's bad. And wonderfully, you won't be tilting at windmills to guess which sources a publisher utilizes.

An ongoing active relationship makes this possible. It proves continued ownership and existence to the publisher. It keeps the "hide" or "redirect" flag active, so the duplicates don't surface. The publisher trusts the information as current, represented by an authorized agent, and accurate.

It prevents duplicates at the most important level—the view on an actual publisher's site.

It's impossible to chase down all the bad listing sources out there, one-by-one. Even if you do, the merge process is equally as guilty of creating dupes as a bad source.

A better way is if publishers (like the aggregators) implement an accessible overlay hide architecture—so all those annoying dupes never appear at the surface.

Best SEO Practices for Publishers

Here's how I recommend publishers deal with dupes for maximum SEO benefit:

1. 301 Redirect Dupes To The Canonical Listing

If you can map the duplicates to a single listing URL, then 301 redirect (aka "permanently redirect") the duplicate URLs to the "canonical" URL for that listing. When the search engines crawl the dupe URLs, the redirect signals the search engines to merge them into the target URL and pass all of the SEO value to that URL. This is the most effective way to harvest as much of the "PageRank" as possible. One of the downsides of this technique is that it might take a long time for the search engines to hit the redirected URLs. To speed things up, consider creating a html sitemap of the redirected dupe URLs and linking to it from the footer or from another sitemap page. This will help the search engines find these URLs quickly.

2. Or Use Canonical Tags

If for some reason you do not wish to implement a redirect, then the next best solution is to link the dupe URL with the canonical URL via a canonical tag. So if you want to merge **site.com/dupe** with **site.com/not-dupe** then add the following to the `<head>` section of the dupe URL:

```
<link rel="canonical"
href="http://www.site.com/not-dupe">
```

As with the redirect strategy, to speed things up, create a html sitemap as described above.

3. Or 404 Them

If you can't do a redirect or a canonical tag, then the next best option is to serve a 404 response code on the dupe URLs and remove them from the UI. In my opinion, this is not a particularly great option as it does nothing to preserve SEO value, but at least it eventually gets the dupe out of the system.

Best Practices for Marketers

Before you go off and get rid of all of your dupes, here are a few things to consider:

1. Check Your Dupes' Rankings Before you De-Dupe

Before I get rid of a dupe listing, I check to see which listings are ranking on various publisher sites in Google for both brand queries and high value commercial queries. If the listing ranks in the top 50-100 results for these, then I'll consider using them as the "canonical" listing. In cases where you have exact-match dupes, it may be difficult to determine which listing you are seeing in the SERPs. In some cases, you may be able to map the listing's ID from the different data aggregators to the publisher, but typically this is hard to figure out.

2 Check Your Listings Regularly

As mentioned, just because you kill your dupes does not mean that they won't come back to haunt you. It's a good idea to build a routine where you regularly check the data aggregators and various publishers to make sure the dupes have not come back. At the least set up Google Alerts for your business name and any distinguishing data that was in each dupe (e.g. a wrong phone number) and it might flag when a new dupe pops up.

3. Check Referral Traffic From Publisher Sites

Use your analytics to determine which, if any, profiles are sending traffic to your site from the various publisher sites where you are listed. The profiles that are sending you the most traffic are probably worth keeping.

4. Check Your Reviews

If the dupes have been around for a while, each of them may have built up some customer reviews on Google, Yelp, etc. In general, you'll probably want to try to make the listing with the best reviews (i.e. most, most positive, most recent, etc.) the canonical listing. Contact Google+ Local Support to get their help in merging the dupes into the listing you want to keep.

5. When In Doubt Get Professional Help

As you can see, dupe suppression is a complex thing. I always say "anyone can unclog a toilet, but if you want it done right, you may want to call a plumber". Same thing with dupes.

Conclusion

- Duplicate listing suppression is a tremendous pain point for local marketers. It has become increasingly important, increasingly complicated and increasingly expensive to deal with at the same time.
- Fixing dupes at the data aggregators does not necessarily fix publisher-created dupes. Nor does it necessarily fix the dupes at the data aggregators.
- Fixing dupes permanently at the publisher level fixes both aggregator- and publisher-created dupes.

One Madison Avenue, 5th Floor, New York, NY 10010