MassGIS Report on Creation of an Address Point Datalayer

FGDC Cap Grant G10AC00206

Submitted November 15, 2012

Christian Jacqz, MassGIS
Information Technology Division
Commonwealth of Massachusetts

1. Project Narrative

1.1. Summary of Project Activities and Key Technical Elements

This project involved the creation and validation of an address point database for a number of communities in Massachusetts A draft address point database was created by integrating several spatial data infrastructure (SDI) layers, specifically orthophoto, LiDAR, roads, tax parcels and roof outlines. Then, using a tablet application developed for this project, errors and questions regarding the draft layer were resolved in the field. The address point database consists of the point locations themselves, which in general lie within structure outlines, and the address records from various sources which were parsed and standardized and loaded into a master address lookup table. Once the geography (the points) and the tabular data (the addresses) were in place, building the initial version of a point address database involved linking the points and the address records so that every address had a geographic location. We used a normalized model, which means that one point location could link to many address records.

The technical approach included several key elements that represent what we believe to be best practices that will be incorporated into a statewide project. The purpose of this report is primarily to cover these key technical elements in enough detail so that others can evaluate and possibly adopt the approach we took.

1. Field layout for numbered address data based on the recently issued Federal Geographic Data Committee (FGDC) standard and for sub-address data based on a profile of the FGDC standard as implemented in the draft Civic Data Location Exchange Format (CLDXF) by the National Emergency Numbering Association (NENA).

2. Use of a flexible and powerful parsing engine in the Python programming language to bring street names to the FGDC standard and extensive coding in Python to standardize other components of thoroughfare addresses as found in local datasets.

3. Development of a relational data model for address points and tabular listings which builds on the available SDI data for the state. This data model replaces the flat-file model currently in use by many municipalities in Massachusetts.

4. Deployment of a browser-based mobile GIS data capture application built on HTML5. Since forthcoming phone and tablet hardware and operating systems are expected to continually improve support for HTML5 geolocation, map display and data management capabilities, it is our belief that a browser-based solution represents the most viable and portable approach for mobile GIS.

5. Use of LiDAR data from the New England ARRA project and from FEMA and other agencies by the vendor interpreting building outlines and by our program in performing QA of those outlines. We were not able to reliably derive structure points from the LiDAR directly, so we elected to go ahead and interpret structure outlines manually from orthophoto, but the LiDAR played a very important role in

the quality assurance of the building outlines, which we estimate improved the accuracy from 99.5% to 99.8% (cut the error rate in half).

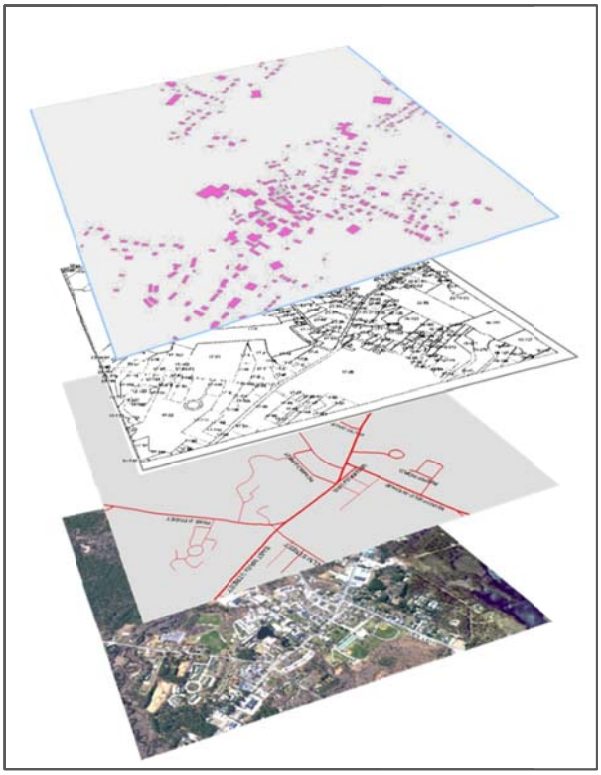
One thing we looked at was how such projects in Massachusetts could support the construction of a Key Facilities layer based on the Homeland Security Infrastructure Program (HSIP) sectors and sub-layers as detailed in the Homeland Security Infrastructure Tiger Team Report, Version 1.1, September 2002, Appendix D – National Critical Infrastructure and Urban Area Minimum Essential Data Sets. There were two approaches to consider. First, we created a lookup from the statewide listing of use codes published by the Mass. Department of Revenue (DOR) for local assessor use to the FCode domains as presented in the Best Practices Data Model – Structures (2006). That lookup is attached to this report as a spreadsheet. But since assessors are focused on setting a fair and equitable market value of property for taxation purposes, the specific type of business activity may not be a concern for them as opposed to the general characteristics of the building and thus the assessor use codes do not offer all the level of detail required in the National Map domains for commercial, retail and industrial properties. Even less detail is provided for non-taxable properties, this different government agencies which may be in different sectors are grouped together by the state DOR. Our recommendation is to pursue a second approach, which is to geocode, using the point address dataset being generated by this project, from address sources other than the local tax list. The best sources we have found for the commercial, industrial and retail domains in the Structures Data Model were commercially available lists of businesses, such as Dun and Bradstreet or InfoUSA, which include the NAICS code. The NAICS codes can then be mapped to the Structures Data Model to provide the full range of FCodes for those sectors. Categorizing the full range of government facilities, on the other hand, is best done by matching data from facility management systems for federal and state government. At the Federal level, GSA has a useful inventory, and in Massachusetts, the Department of Capital Asset Management's Capital Asset Management Information System (CAMIS) maintains a similar listing of state facilities. For municipalities, there are a range of sources based on the vertical relationships between state agencies and local departments – Mass Emergency Management has a pretty good listing of public safety and public works sites, whereas the Department of Education compiles school locations, Department of Environmental Protection has locations of water and wastewater treatment plants and so on. The single biggest challenge is the mapping of the variety of infrastructure covered in the Transportation domain – the state Department of Transportation has many but not all of these layers mapped in separate inventories of bridges, airports, rail facilities and so on.

As noted, the key resource for this work is the point address dataset. The address data compilation for which this project served as a pilot was initiated at the state level primarily to meet public safety needs, but as is often the case with GIS projects, there are many ancillary benefits accruing to project partners and participants. We used the project funding to directly support the involvement of regional agencies, who are enthusiastic about a long-term partnership on this activity. Other state agencies and local officials participated in the project on a volunteer basis because of their interest in the project outcomes. Since address

data are so widely used it is not surprising that this kind of project would motivate the participation of regional and local partners. However, a key lesson learned is that the organizational and institutional challenges are as least as great as the technical ones. Volunteer energy only goes so far, and in Massachusetts at least, the issue of "unfunded mandates" is a very sensitive one. Making the case that geographic data collection of addresses should be added to all the other responsibilities of local officials requires substantial persuasion, and one lesson learned is that we will have to allocate more staffing and resources to that effort. We need to address the lack of an overall mandate for integrating local and state data, lack of understanding and support for project goals, and very uneven technical capacity among project participants at the local level. Massachusetts has no counties, and so the responsibility for data collection and maintenance falls to the municipalities, which often lack any technical infrastructure.

1.2. Background - Strategic Plan for Mass. Spatial Data Infrastructure (SDI)

In June of 2007, MassGIS published a *Strategic Plan for Massachusetts' Spatial Data Infrastructure.* The plan, which was funded by a USGS Cooperative Assistance Award (CAP grant), laid out the priorities for development of four key NSDI data layers. The plan reflected a statewide process of information gathering, stakeholder workshops, standards review and budget analysis which generated widespread support for the SDI agenda. The layers identified in the plan and featured in the graphic at left, were orthophoto and elevation data, roads, tax parcels and structures.



The mapping of tax parcels, managed in Massachusetts at the municipal rather than at the county level, was identified in the Strategic Plan as the single largest data gap. According to the report, there was a serious "digital divide" between those communities that had automated their parcel mapping and those that had not. Another issue for parcel mapping was the lack of mandate - a minority of communities had voluntarily adopted the state standard for parcel data but there was, and still is, no requirement for communities to do this, with the result that mapping varied wildly in quality and format. The report recommended a multi-year capital program to automate and/or upgrade existing parcel mapping to a new statewide standard designed to

support the typical municipal applications such as assessing as well as state agency and private sector needs.

The Strategic Plan did not propose a statewide mapping of structures, but rather focused on the need for more detailed and geographically accurate mapping of critical infrastructure. However, the need for structure mapping was implicit in the identification of address points as another critical data gap, as described in the following section:

> ***Statewide address locations are required:*** *Myriad activities and responsibilities of state government are linked to a physical address, ranging from environmental permitting to day care licensing to police and fire response. Using GIS technology, it is possible to geocode or automate the mapping of address locations. This can help agencies deliver services more rapidly and effectively, analyze patterns and identify constraints or opportunities in combination with other layers of GIS information. However, this all depends on having high quality geocoding data. Currently available data sets were identified as inadequate and in need of improvement, particularly in a public safety context.*

In another section, the report identified the State 911 program as a major beneficiary of improved geocoding based on parcel addresses and as a likely source of funding for such work.

1.3. Implementation of SDI for Massachusetts – Public Safety Requirements

Since the Strategic Plan report was issued, the statewide GIS program has been moved from the environmental agency to the central IT agency and has undertaken several major initiatives recommended in the Strategic Plan. Specifically, the business case to automate or upgrade all the municipal tax parcel mapping into a statewide layer was further detailed in a follow-on Parcel Business Plan, also funded by a CAP grant, and was initiated with cooperative funding from information technology bonds and State 911 as suggested in the original Strategic Plan. The 911 involvement was based on two public safety requirements faced by jurisdictions all across the country as a result of the impending move to the next generation of 911 technology, called NG-9-1-1. The deployment of NG-9-1-1 requires that each call be linked to a point location as part of the call routing process, and that emergency service zones representing areas of responsibility for individual call centers be digitally mapped so that the call can be routed using GIS point-in-polygon overlay rather than a tabular lookup as in current systems. The mapping of parcels provides an initial address point location which is an improvement over linear geocoding and thus supports the first requirement. However, another compelling argument for the parcel mapping to support 911 is that the mapping of emergency service zones in Massachusetts will be based on aggregating parcel boundaries.

The lack of standards for parcel mapping at the local level meant that each of the 351 communities in the state presented a unique challenge, not only with regard to mapping accuracy but also for standardization of the address information contained in the assessor tax records. Standardization of addressing is needed operationally to support matching the input (caller address) to the geography (points with addresses derived from parcel layer). For 911, we did not consider using commercial geocoding software with a set “tolerance” for

mismatches because that would introduce unpredictable errors – when we use the term "match" we mean a geocode score of 100%, that is an exact string-equality match. Our report highlights the importance of the Federal standard for addresses in this context; although of course any geocoding application will require some degree of standardization and quality assurance for the GIS and other input data being used, the particular requirement of 911 for match rates of 99% or higher means that much more effort needs to go into pre-processing and standardizing the address data. The parcel mapping work was contracted out to multiple vendors over a three year period, with every effort made to leverage existing relationships between communities and vendors. It will be completed by March of 2013. Address standardization was done in-house and will be complete by June of 2013.

The proposal in the Strategic Plan to map critical infrastructure at the level of individual structures evolved into a proposal to map *all* structures, because of the public safety requirement cited above to attach a point location to every call in order to support 911 routing and dispatch. While parcels provide an initial site address there are two reasons to seek the additional detail from mapping structures. First, in almost all cases, emergency services are responding to a structure and knowing where on a large parcel of land development has occurred, will help responders find the correct location more quickly. Second, on some large parcels there may be many structures which need to be identified at a greater level of detail in order to support timely dispatch (e.g. college campus, trailer park). The development of a point address dataset based on structures allows for refinement of the address to include building names and other identifying information. In sum, attaching address information to structures, along with mapping the emergency service zones, provides the complete GIS solution for NG-9-1-1.

As noted above, the Strategic Plan called out the challenge of working with local data, and to a large extent that issue remains. The structural difficulties associated with collecting and managing data from small municipalities which have little or no technical capacity, some with populations of only a few hundred, have not gone away. The goal of data flowing from the "bottom up," which has been a mantra of the national spatial data infrastructure program since its inception, may not be realistic in such situations. The model developed in Massachusetts reflects the particular circumstance of the state in that a great deal of the work for the initial development of the SDI has been centralized at a state level with the local function limited in some cases to review.

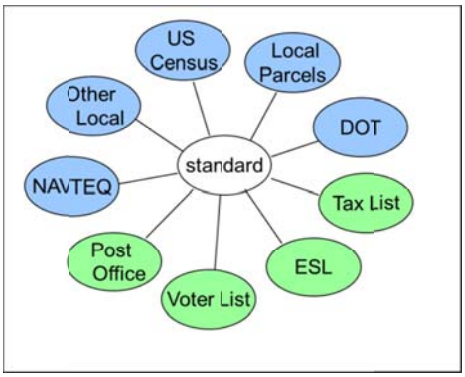
Nonetheless, the responsibility for quality assurance and maintenance of some GIS layers clearly must reside at the local level for both jurisdictional and logistical reasons. The report details how our technical approach deals with this reality by pushing out a "draft" version of the data and using web-based access for subsequent review and field data collection by regional and local staff. Our report clearly recognizes that the challenge of this approach is as much about motivation and organization as it is about technical design. We believe that the role of regional staff in providing both the missing technical capacity and the local presence is key.

1.4. Overview of key project elements

Besides highlighting the absolutely essential role of standards for parcel mapping and for addressing, this project also showed the importance of the synergy between NSDI layers and of the development of a point address data model which formally links those layers to maximize the benefits of an integrated geo-processing environment. One kind of synergy involves using geographic relationships to validate tabular data. In the most basic kind of example, the name associated with the street segment on which a parcel fronts is expected to match the name in the address field. Only a rigorous standardization can support this kind of matching in an automated fashion, but once it is in place, the benefits of finding any discrepancies in street naming are significant. A more complex kind of synergy involves the use of linkages based on overlay. An example here is to require a numbered address for every parcel where there is a structure, indicating some level of development. Other kinds of overlay involve image classification, such as investigating changes detected through comparison of images to ensure that there are address points everywhere automated change detection suggests, and field inspection confirms, that there has been development on a previously undeveloped site.

Implementation of FGDC standard for street names

As discussed above, there are several important reasons to implement the FGDC address standard, both in the public safety context and generally for all government agencies managing address data. First, why implement a standard at all? Any agency wishing to improve the quality of geocoding should consider implementing a standard. In particular, a standard supports efficient matching between input lists and geography in the initial construction of a point address dataset. Because no single list being matched is likely to be comprehensive, the standard supports identifying those records that are common and those that are not between different lists. But getting a single authoritative list of addresses is only half the battle – then that list has to be matched to a variety of geographies such as linear address ranges on street segments or addresses associated with parcels from assessor maps. The standard supports that matching in a transparent fashion, unlike geocoding which has a "black box" component that can lead to errors that are hard to find. Finally, in integrating any new tabular source of addresses, or a new iteration of an existing table, the first requirement is to determine if there are any new addresses. Again, this can be done by database matching, but only if there is a lookup from all the various forms of the address to the standardized form. The graphic above illustrates this.



ata

Another important purpose of the standard is to parse addresses so that the numbered address portion (e.g. *100 Main Street)* can be separated out from the sub-address portion containing additional detail, often called "location" information (e.g. *Rear* or *Unit B* or *Marshall Building).* Processing addresses will often involve filtering at different levels of the

address hierarchy – so that complete address records containing all the detail down to the unit level can be grouped by street name, or by range of numbers, or by numbered address, or by building. An address standard with a hierarchical structure supports this kind of query very well.

The choice of the FGDC standard for our address processing is based on two distinct benefits it provides – first, its schema for the elements of an address works better than any other to handle the variety of street names and address formats that are out there and second, it supports the need for hierarchical queries against some or all of the address information. By comparison, USPS Pub. 28 has limited flexibility in cross-walking to other datasets, doesn't handle a number of common element sequences, has no structure for sub-address elements and has limited domains for numbered address elements. No other standard that we looked at handled sub-address information as well.

Prior to this project, the FGDC standard was used to parse several million input address records to generate a lookup table of standardized street names for statewide use, called the Master Street Lookup or MSTR. Inputs included street names from our current street file (described later), TIGER, Mass Department of Transportation (MADOT), property tax lists and the 911 Master Street Address Guide or MSAG. The MSAG is critical because it lists address ranges and street names for 911 data validation purposes and is supposed to be authoritative. The MSAG is centrally managed by Verizon (the dominant land-line carrier in MA) with support from local 911 data managers. If Verizon gets a request for a new phone number whose site address number does not fall within existing ranges on a given street and community, or where the street name does not exactly match an existing street name within the given community, the MSAG file has to be edited by local 911 data managers to accommodate the new address information before the new phone number can be approved. MassGIS obtains the MSAG file on a regular basis from the State 911 program and runs it against the MSTR to keep the two files in sync.

Address Number – Prefix, Number, Suffix
- prefix "B18", "Milepost 12.2"
- number "247"
- suffix "12A", "14 ½"
- ranges "12-14"

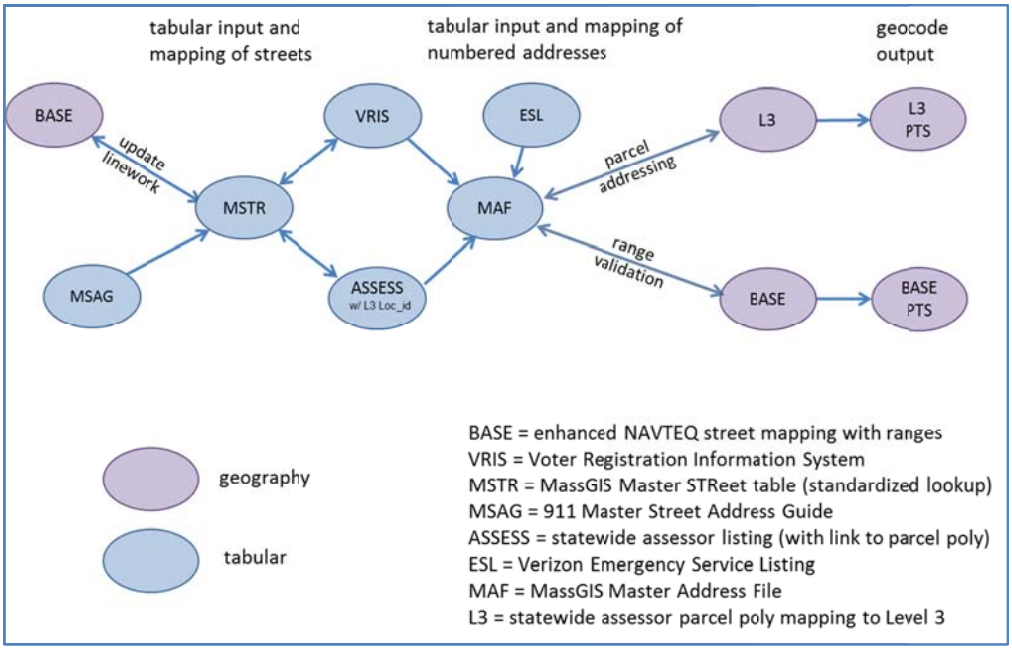
Street Name – FGDC says fully spelled out
- pre-mod "Old North Coach Road", "Upper Hampden Road"
- pre-dir "South Main"
- pre-type "Avenue A"
- street-name "Broadway"
- post-type "Market Street"
- post-dir "Washington Street South"
- post-mod "Charles Street Place", "Chatham Street Extension"

Sub-address (postal secondary location)
- Type, name pairs as needed
- Examples: "West Campus, Building A Rear"
  type1="Area", name1= "West Campus"
  type2="Building", name2="Building A"
  type3="Position", name3="Rear"

The MSTR list was used in this project to validate all street names and to provide a pick list for street name data entry. The names of any new streets from any input source were captured, standardized and added to the statewide list as the project went forward.

Workflow – streets and parcels

As described in the overview, one of the challenges in construing a point address file is that the different workflows involved integrate data products that are managed at the state level, sometimes with data entry by local officials, with other data products that are compiled and managed at the local level but aggregated to the state level for the purposes of this project. This particular workflow integrates the streets and parcel layers with the Master Street



ile

Address Guide or MSAG and with other sources of address information to construct a Master Address File ("MAF" in the diagram). The streets are managed at the state level, in a dataset ("BASE"), which is a derivative of a licensed commercial street map from NAVTEQ.

The streets dataset is licensed for all levels of government in the Commonwealth and is currently being used to linearly geocode input addresses for 911 and for a variety of other applications. It is updated quarterly. The parcel mapping, on the other hand, is held at the local level and MassGIS is in the process of automating parcel data from hard copy, or enhancing and standardizing existing GIS data, for 350 discrete jurisdictions, including the towns for this pilot. What's being delivered on a one-time basis is a data extract from the tax file ("ASSESS" in the diagram) linked to parcel polygons ("L3") via a unique location ID. Thus, the workflow outlined above involves processing many separate files which originate externally to the program. Since there is as yet no mandated annual submission of parcel data, *the key to working with all these datasets is to maintain a link back to the original data source, whether it be NAVTEQ or a local assessor file, in order to be able to track changes and capture new information.*

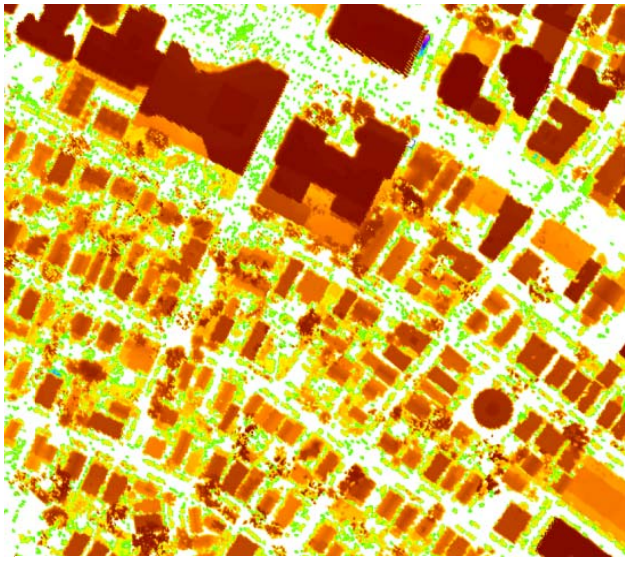
The workflow diagram illustrates the importance of the standardization process. We enforce consistency between the street names in the different datasets using a table called Master

Street Lookup ("MSTR" in the diagram) which stores all the original forms of the street name along with the authoritative name. Every street name in the MSAG is linked to a standardized street name in MSTR, and then that lookup is checked against BASE to make sure that street is mapped. (see sample records above). With MSTR, MSAG and BASE all in sync, the tax list ("ASSESS") is processed to ensure that all the street names listed for parcel site addresses can be matched. Any streets that are not already mapped can be located using the right-of-way (ROW) shown on the parcel geography linked to the tax records and mapped heads-up using current ortho, or if the street was built after the date of the imagery a "stick-figure" street segment can be stubbed in for later field collection using the parcel ROW as a guide. For streets that are already mapped, the street map can be used to ensure consistency between the parcel street names and those in BASE. The geographic proximity between the parcel and the street is used in an automated fashion to flag situations where the standardized street name does not exactly match the standardized parcel street name and these situations are manually reviewed with current orthophoto. It's extremely important to avoid making any assumptions about whether or not street names that differ actually refer to the same feature, and to manually review the orthophoto, the streets and the parcels in all such situations to ensure consistency. For example, if the type of the parcel street name doesn't match but the base name does, there may be a cul-de-sac or small shared drive off another street with the same name – e.g. "Marshall Court" may be a little cul-de-sac off "Marshall Street."

LOCATION_TYPE	FULL_STR	FULL_STR_STD
TIGER_09	Garden Street Ct	GARDEN STREET COURT
MGBASE_1007	GARDEN ST CT	GARDEN STREET COURT
NT_09Q3	GARDEN STREET CT	GARDEN STREET COURT
MSAG_090512	GARDEN ST CT	GARDEN STREET COURT
WG_2009	Garden Street Court	GARDEN STREET COURT
MSAG_090512	E Spring St RR Xng	EAST SPRING STREET RAILROAD CROSSING

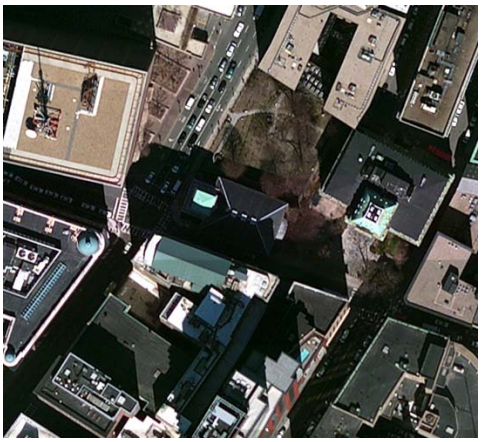
Once the street names in the tax list have been checked and standardized, the next stage of scrubbing is now focused on the street numbers in the tax list – any additional content contained in the street number field needs to be cleaned out and put into a sub-address field. For example, the assessor may have unit information in the number field, e.g. "10-A Market Road," "10-B Market Road," etc. to indicate units "A," "B" and so on within a single structure addressed at "10 Market Road." Following the FGDC standard, the numbered address should be "10" and the unit information should go into a sub-address field. After the numbers have been scrubbed to correct the many similar sorts of idiosyncrasies, the addresses in the tax list are added to the Master Address File ("MAF"), as shown in the flow diagram above, along with the unique geographic identifier for the parcel polygon which links the tax record to geography. The end result of the process above is the assignment of a standardized address, which has been validated against the street map, to each parcel. The numbered address now becomes the source for validation of the address ranges contained in BASE and for the generation of address points through overlay with the parcels as described in the workflow for structures below.

Workflow - Ortho, Structures and LiDAR

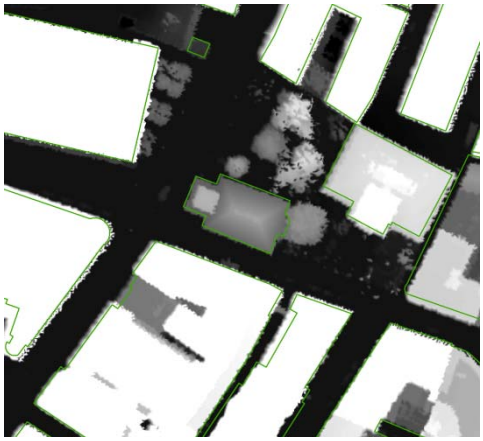


After the need for point addressing to support 911 call routing and dispatch became clear, we attempted to create a set of structure centroids using LiDAR from the ARRA, FEMA and other projects. We did not have the tools to use the classified .LAS point files directly (such tools are now available in ArcGIS 10.1) and instead attempted the classification of a normalized digital surface model or NDSM, which is derived as the difference between the bare earth elevation model and the digital surface model (including structures, trees and other above ground features.) The major difficulty is in distinguishing trees from structures – which is easy to do based on visual characteristics but surprisingly difficult to do with image classification. Despite applying considerable expertise to the problem in the eCognition and ERDAS Imagine environments, the error rate still remained too high for practical application.

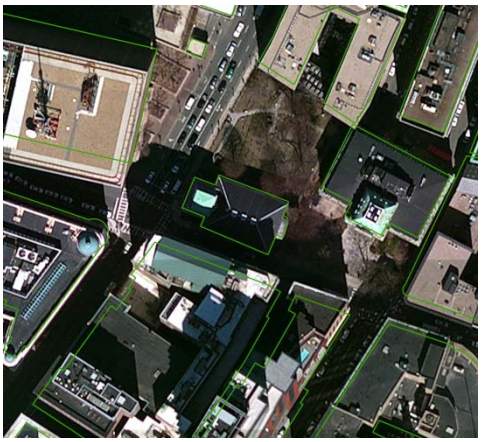
The fall-back approach was to procure structure outlines from a vendor who interpreted them from orthophoto in a specialized CAD-like production environment. The cost of doing this was about $0.08 per structure for 2+ million structures. As described above, the spec for the structure outlines was pretty exacting and LiDAR ended up playing a role in the production and the quality assurance. We required that the sum of the error rates of omission and commission be less than 0.5%. The imagery used was from Digital Globe (both Standard and High-value products) and we were pushing it to the limit to support this level of accuracy. However, when the LiDAR was used to supplement it, the vendor was able to drive down the combined error rate to about 0. 2%. The outline product is shown at left. The use of LiDAR data is shown below.



Ortho with building at center of image in shadow



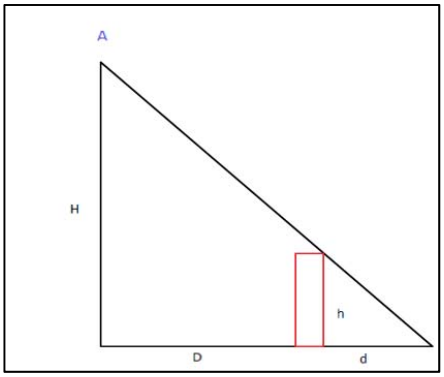
LiDAR of same building showing extent



Ortho with building outline

Roofprint shifted using LiDAR

The first image shows a problematic situation in which a number of building outlines are partly in shadow. The most common problem, in urban environments, was a narrow alley between two buildings, obscured such that it could not be determined whether there was one building or two. LiDAR was used to complement the ortho interpretation and the QA, as shown in the sequence above. Another way in which LiDAR was used to enhance the footprints was fairly novel – what might be called "poor man's rectification." As shown in the last image in the sequence above, roofprints in an ortho are shifted from the "true" ground position of the building footprint because of building lean. The geometry of building lean is shown in the graphic at right. Given the height of the camera used for the ortho, about 5,000 meters, the shift was minor (<5 meters at the edge of the frame) for two- or three-story buildings. But for taller buildings, it was problematic. We were able to use the LiDAR and the NDSM generated previously to estimate



D= Distance from frame center
H = Camera height
h = building height
d = displacement

the height of the building, h, and mapping of the roofprint relative to the center point of the image to estimate D, so since D >> d, using similar triangles and solving for d, d ~ D/H * h, and we could calculate the magnitude of displacement. The direction of displacement was simply away from the center point, so a python script allowed us to complete the approximate correction of the roofprint shift for all areas where we had LiDAR. (note: this is *not* photogrammetry, but it mostly solves the problem of roofprints overhanging parcel boundaries because of building lean, which otherwise creates "slivers" in the overlay of buildings and parcels. Remember, we are developing only what we need to create address points in a cost-effective manner. Planimetrics would have been cost-prohibitive.

Relational Data Model

The most significant and useful part of this project was the focus on working out a data model for address points that could meet the needs of 911 and other address data users. The key findings which informed the design of our model are listed below:

1. The most cost-effective source of address location information, initially, is the parcel mapping maintained by local assessors and then compiled on an ortho base to a standard for geographic accuracy such that parcel boundaries align correctly with visible features (e.g. structures fall into the "correct" parcels). As discussed above, the other SDI layers provide checks and validation for the address locations in the parcel layer. For example, a parcel with a developed use code, or one where a structure is visible on the photo, should have an address. The parcel address would be expected to have the same street name as at least one street fronting the parcel. There should be agreement between ranges and parity for street segments and parcel address numbers. The data model has to support all these kinds of relationships, both spatial and tabular.

2. In building an address list with available information, at least from a statewide perspective in dealing with many small municipalities (and no counties, as noted above, we don't have them!) multiple sources must be conflated, because the traditional source of addresses for 911 in the Emergency Service List of land-lines is becoming less and less comprehensive as mobile technology displaces land-line use (this is commonly described as land-line customers "cutting the cord"). This need to conflate addresses drove our use of the FGDC standard as discussed above and also determined our approach to the data model. The pragmatic approach to building a comprehensive list of addresses takes all address lists that are available, standardizes them and compares them to come up with a unique list. We found that a single "look-up" table with all various address records included in it, one which preserved the link to the original source using IDs and metadata, was the easiest way to effectuate the reconciliation of different source data. This was primarily because we had neither the time nor the authority to "correct" the original source. Even at the local level, such authority over other departments may not be available to an addressing project in one department. This step provides tabular validation of the parcel address info.

3. Back on the geographic side, we concluded that it was best to use "real" physical features, i.e. buildings to represent address locations. We also determined that it was better to assign a single address to a cluster of buildings at a site than to attempt to determine from the ortho which was the "primary" structure. The simplest case is a single-family lot with several structures, one structure that in context looks like the residence (based on clues like pavement and landscaping) and another that is somewhat larger and might be a barn or a garage. First of all, it's a lot of work to interpret all those clues and it has to be done manually. More importantly, you could be wrong. To assign the address to a single structure risks missing the in-law apartment, or the business, or some other use in another structure to which emergency personnel might have to respond. We determined early on to use all the building centroids in the initial draft of the address points.

4. On the other hand, there will be many situations in which multiple addresses are associated with a cluster of buildings, such as a condo complex where there are several buildings each with multiple units. We need to support the use case where a list of addresses, not just one, is linked to a collection of buildings without specific assignment to any single building. We were very cognizant from our management of parcel data that a many-to-many relationship between structures and addresses can be confusing to end users and difficult to manage in a GIS environment. Many common GIS functions like query, symbolize or identify become more complicated with a many-to-many model.

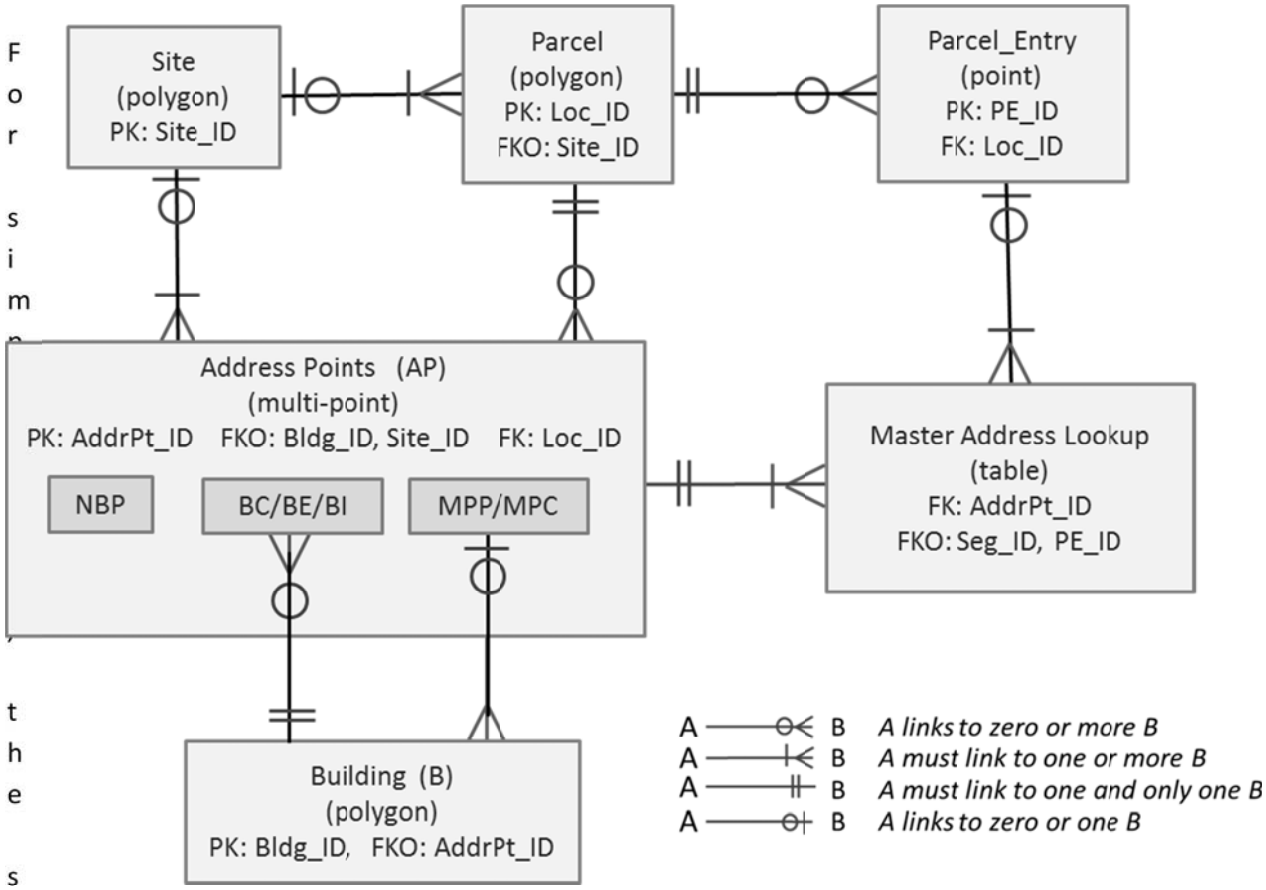
5. Also, there will be a significant fraction of parcels for which the default address will not suffice, and more detailed address information at either the numbered address level or at the sub-address level will be required. These include campuses, condo complexes, hospitals, large industrial facilities and many other commercial, institutional or residential "sites" that cover a large area and have many buildings (often fronting on many different streets.) The data model has to support adding information with field data collection which enhances the geographic accuracy of the address point associated with specific address records grouped together at a site.

6. We concluded that the best approach to deal with these requirements is to group the structure points into a single feature using the multi-point feature class in ESRI software by dissolving on the site address field. While this is a proprietary solution, we felt that it was admissible given the overwhelming dominance of the GIS market by ESRI. "Dissolving" address points representing structures into "multi-part" features solves the problem of many-to-many and avoids the redundancy of repeating the address attribute for each point. Normalizing the relationship, by establishing a one-to-many relationship between (potentially multi-part) address points and address records, is better database practice than the one-to-one approach of "stacked" or "shotgun" points. Even in the simplest case where many address records can be assigned to just one building, the approach of stacking points can be very awkward. Consider the example of an apartment building with potentially hundreds of unit address records. Labeling, querying, and overlay operations on such a stack of points become cumbersome. In

particular, editing operations are confusing and difficult. But the real decisive factor is that the stacked approach can't handle the case where multiple addresses are associated with multiple structures – this may only be a temporary situation, where the points will be disaggregated as described above, but it still is a set of records and points that need to be managed. Otherwise, in the stacked point scenario, addresses in a many-to-many situation would have to be linked to an arbitrary point representing a cluster of buildings until field work was done to disaggregate them.

The decision to manage GIS address points and address records separately in a normalized address model is not necessarily consistent with how other jurisdictions are approaching this problem. The good news is that even in a many-to-many situation, a spatial view, with the locations of multi-point centroids repeated for all address records at a given site, can provide the "flat file" that may be needed for systems and standards yet to be finalized.

Below is a diagram of the data model and following that, illustrations of specific cases from simplest to most complex.



Note that **streets** and the **master street list** have been left out of the diagram above to show just the possible configurations of an address point feature. The relationships in the diagram

should make some sense given the findings above, and the following additional discussion will explicate the notation.

In the above diagram, an **address point** is linked to one or more address records. A structure point could be linked to no address records, e.g. a deer checking station in a state forest, but by definition, every address point will have an address. An address point can be a **non-building point (NBP)**, such as a playing field or a parking lot. It could also have just one part, which could be a **building centroid (BC), building entry (BE), or building interior (BI) point.** Or it could be **a multi-part feature, made up of more than one building centroid (MPP).** Note that technically, all address points are included in the multi-part point feature class, but some have just one part. The label information for a multi-part point feature will be carried by the **centroid of the multi-part point (MPC)** which is created to provide a unique ID and because individual address records will need to be summarized for labeling purposes in one address record (e.g. a range of address numbers or range of units rather than individual numbered addresses or sub-addresses). As noted above, the centroid of the multi-point may be used to carry the address information in an export to any required flat-file format.

A **building polygon** where the building has at least one address associated with it will always be linked to one or more **address records** via one or more **address points** which lie within the building envelope. In one possible case, the building centroid will be part of an address point feature, either by itself or joined together with other building centroids in a multi-part feature (BC or MPP). In another case, address points will be located at individual building entry points (BE) each one of which will be linked to a distinct address. Entry points must be geographically accurate, that is they must be located within the building outline near the actual entry. In a third possibility, which is the most detailed kind of representation, individual address locations within a building, such as stores within a mall, will be mapped. These are labeled BI for building interior in the diagram.

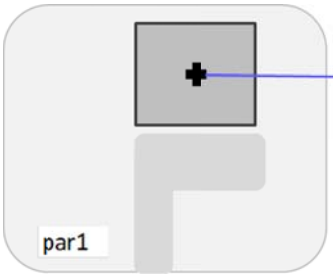
A **parcel** may or may not contain any address points. In some jurisdictions, every parcel has an address, whether or not it is developed. But in others, there might be no address record. A land-locked parcel might not even have the street name listed.

A **site** is our term for an area comprised of one or more parcels, or an arbitrarily drawn polygon which may include parts of one or more parcels, for which we have determined that additional detail is needed for the buildings within the site polygon. This means that within the site there are no multi-part points, or more correctly, that all address points will have just one part. In other words, every building will be identified as a distinct entity rather than being grouped together with others. A parcel can have multiple buildings without becoming a site.

A **parcel entry point** is typically the point where the driveway to access a given address meets the road and is especially useful for dispatch in situations where the end location may not be visible from the road or there may be alternate means of access but one is preferable.

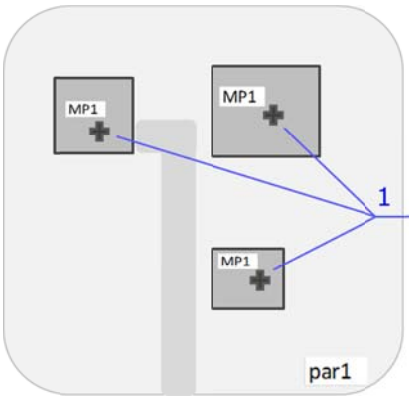
The following diagrams are intended to clarify how the model was implemented.

1. **Simplest case – one parcel, one building, one address- building centroid**



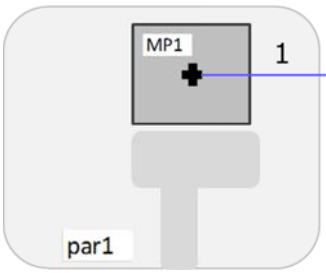
LOC_ID	ADDRPT_ID	FULL_ADDR
par1	MP1	1 CAMDEN WAY

2. **Most common case – several buildings, one address – multi-point**



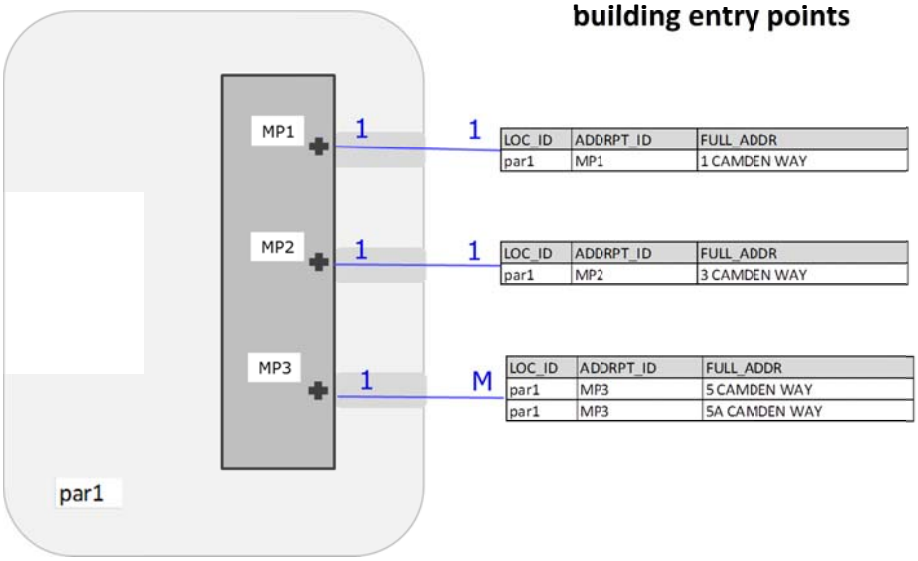
LOC_ID	ADDRPT_ID	FULL_ADDR
par1	MP1	1 CAMDEN WAY

3. **Also common– one building, multiple addresses – building centroid**

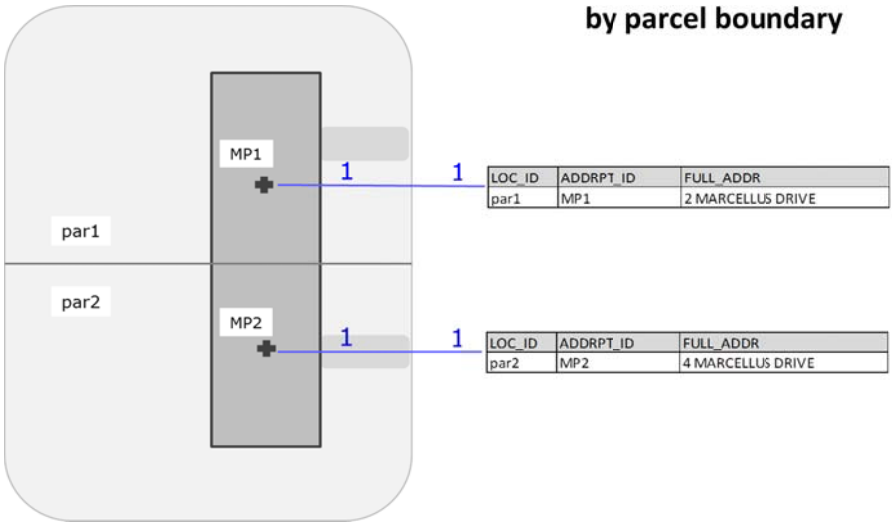


LOC_ID	ADDRPT_ID	FULL_ADDR
par1	MP1	104 NORTON AVENUE #10
par1	MP1	104 NORTON AVENUE #11
par1	MP1	104 NORTON AVENUE #12
par1	MP1	104 NORTON AVENUE #13
par1	MP1	104 NORTON AVENUE #14
par1	MP1	104 NORTON AVENUE #15
par1	MP1	104 NORTON AVENUE #16
par1	MP1	104 NORTON AVENUE #17
par1	MP1	104 NORTON AVENUE #18
par1	MP1	104 NORTON AVENUE #19

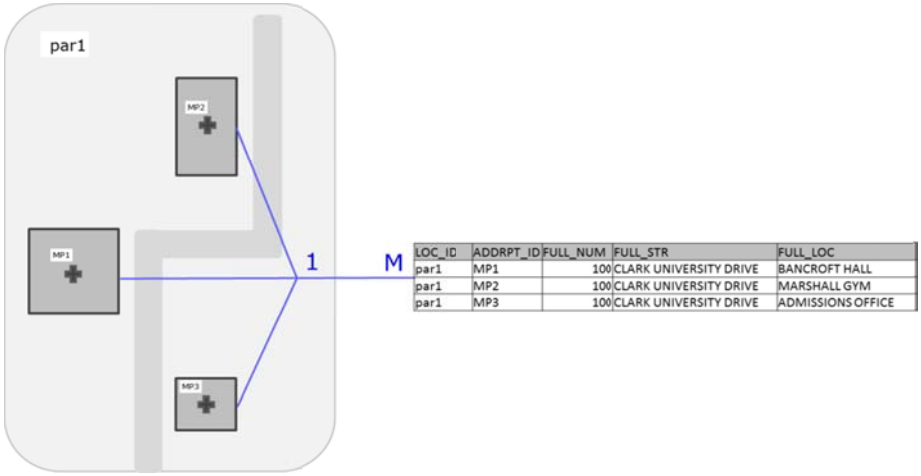
4. Higher geographic precision – one building, three entries, four numbered addresses - building entry points



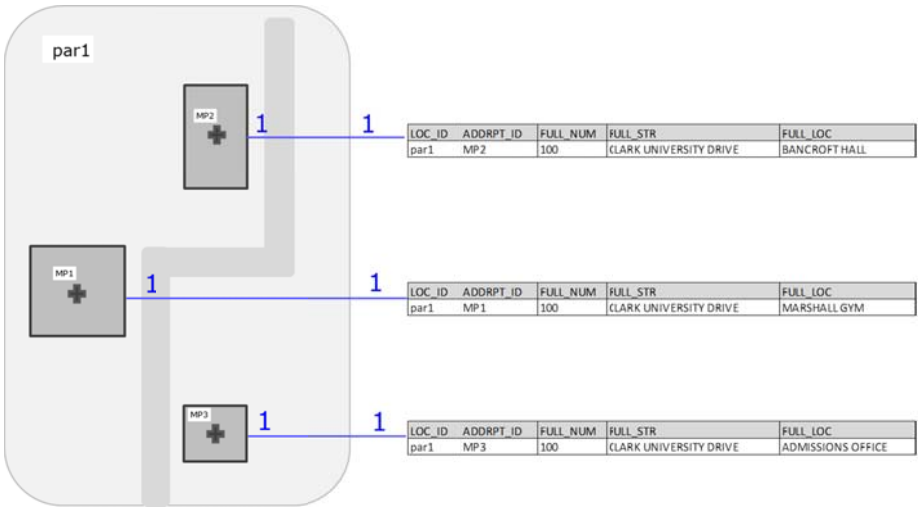
5. Two parcels, one building, two numbered addresses – building centroids based on split by parcel boundary



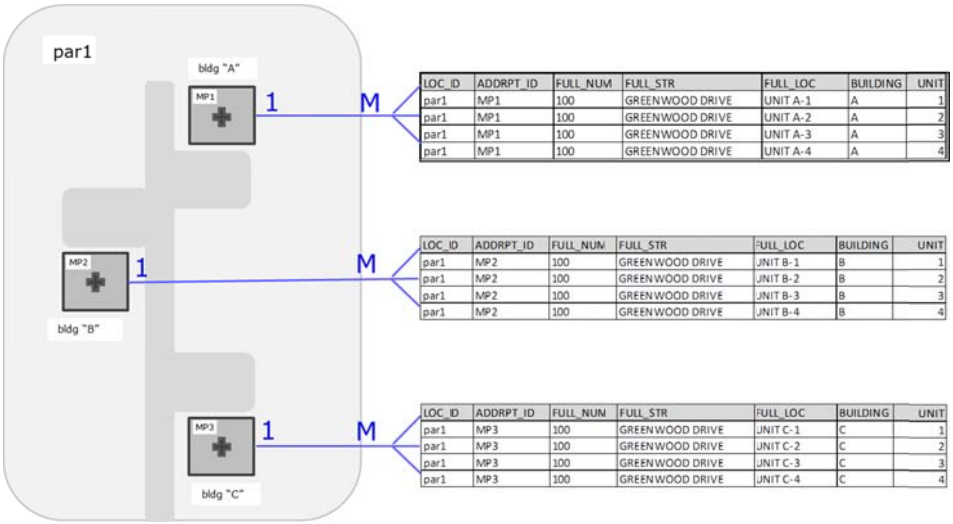
6. Multiple buildings, multiple sub-addresses (may need to be disaggregated) – multi-point



7. Multiple buildings, each building labeled ("exploded" multipoint from 6)



8. Multiple buildings, each with multiple sub-addresses (e.g. condo units)



Field data collection background

The field data collection portion of this project was fairly compressed but extremely valuable and we intend to build on the lessons learned to further invest in the development of tools and methods. The workflow descriptions in earlier sections of this report covered the collection of roads, parcel and building location data. Also described was the construction of a master address lookup from a variety of sources with a standardized format for number and street name. As noted, for "sites" we are standardizing the building name or other location information as well as the numbered address. After the geoprocessing of these inputs is complete, for any given municipality, we end up with a "draft" point data set where the parcel addresses have been standardized and transferred to building points, and one or more sites have been designated as locations where additional detail is needed. What's left over from this process is a set of addresses which do not link to any location, some that link

to a general area but need a more specific location, and some points that do not link to any address. The final phase of this project was to develop a mobile browser-based application to deal with these cases. The application supports upload/download, using Open Geospatial Consortium Web Feature Services and other database protocols, of addresses and point locations to be found or visited. Although the intent is to standardize all addresses to the extent possible, lists of locations to be found can also include free-form descriptions such as "new house at 41 Lincoln St. between North Elm St. and Marlowe Road" or "Administration Building on the Williams College Campus."

Design Objectives for Mobile Application

We had the following objectives relating to the design approach, exclusive of the functional requirements. We wanted the app to be **tablet-based,** and the use of the application in the field confirmed that a tablet form factor, at least a 7" screen was preferable to a smaller screen, given the complexity of the mapping to be displayed and parcticularly the need to zoom and pan while viewing aerial imagery. We also wanted it to be **portable** (non-proprietary), so that we could deploy to any number of different operating systems or devices. Another important goal was to be **user-friendly,** such that we could deploy to municipal staff who were not necessarily technical, with a minimum of training. We wanted the app to be **wireless but also autonomous,** such that if a user lost network coverage they could continue to work. This is often referred to as "detached" editing. Finally, we wanted the app to be **location-aware,** i.e. exploit the GPS capability of the device to filter information presented and to allow the user to see their approximate location displayed.

Technical Choices

Based on the above design objectives, we wanted a fully functional but non-proprietary platform; we found that a browser supporting the full featureset of HTML5 – local data storage, indexed database, forms and the geolocation API (technically a separate module) met this goal. A pure Javascript/HTML solution is the most portable from device to device and avoids any potential issues with vendor-specific technologies not being supported by other vendors (e.g. Apple and Adobe). We also opted to put the business intelligence in the browser code rather than on the server, so that interaction with the server could be limited to standard API calls to the web mapping services. This also supports portability in the sense of allowing the application to be modified and used by others. The caching of both points and addresses supports the editing for momentary disconnects, but we discovered that it was worth the extra memory to cache the imagery basemap tiles as well.

Map Display - The application displays all address points that are considered "correct" in blue, and those which need to be visited in red. Points that have been edited are in green, and those to be deleted in gray. The user is able to zoom and pan and query existing address points including those being collected in real time. The display "follows" the user using the GPS tracking capability. When selected, a point turns yellow, and the associated address, if any, is displayed at the bottom of the screen. The user is able to link the point to an address in the un-located address list, or edit the address provided.

Address List Display – The application displays a list of street names which have unresolved issues. When the user selects one street, they get a list of numbered addresses on that street that need to be found. The user is able to add a point for a given address. They are also able to add an address record and link it back to a point. Ideally, the street and address lists also "follow" the user, in that they are filtered by geographic proximity to the current location using address ranges on the street segments.

Since E-911 response is a key reason for building the point address data layer, maintenance of the address-related datasets requires capturing addresses and locations associated with new development as soon as possible. This means mapping a new road as soon as it is built, and in the case of new structures, creating a geographic point as soon as (a) an address is assigned by the municipality and (b) the physical location of the structure can be determined. For example, if a building permit has been issued and it includes a street address for the construction of a new residence, once a foundation is poured, then it would be possible to visit the site and capture that location.

As part of this project, MassGIS and the Regional Planning Agency partners surveyed towns to record local address workflows. The back–end data maintenance to support the field data capture application includes adding any new address to the master address file when we are notified of its approval at the local level.

Before and during the project, we identified use cases for the mobile field data capture application, from most common to less common. The prototype application that we developed did not in fact handle all of the cases in the list which follows, because we added some as a result of the experience gained during the initial deployment. Here is the complete list:

1) Link one or more selected addresses to a selected MP. This updates the AddrPt_ID of the address record to match the MP.

2) Link one or more selected addresses to a selected part of a MP, thus making the geographic location for those addresses more precise. This creates a new MP and AddrPt_ID from the selected part and updates the address records to match.

3) "Explode" an existing multi-part MP, thus creating a set of MPs with new AddrPt_IDs that need to be linked at some later time to new or existing address records.

4) Create a new address record which remains selected and link it to a selected MP (usually an individual structure) by updating the AddrPt_ID of the address record.

5) "Clone" an address record to create a new address record, edit the cloned record which remains selected and then link it to an existing MP.

6) "Clone" an address record to create a new address record, edit the cloned record (e.g. add unit information) and then link it to a new MP.

7) Create a new MP which remains selected and link it to one or more selected addresses.

8) Create a new MP and one or more new address records and link them.

9) Delete an existing linkage between an MP and an address (note that this will create a situation that requires a new round of editing.) This would be done by deleting the AddrPt_ID.

10) Delete an address that doesn't exist. Any points that are orphaned by this deletion will need to be reviewed.

11) Delete an MP that physically doesn't exist (rare). Any addresses that are orphaned by this deletion will need to be re-linked to another point or themselves deleted.

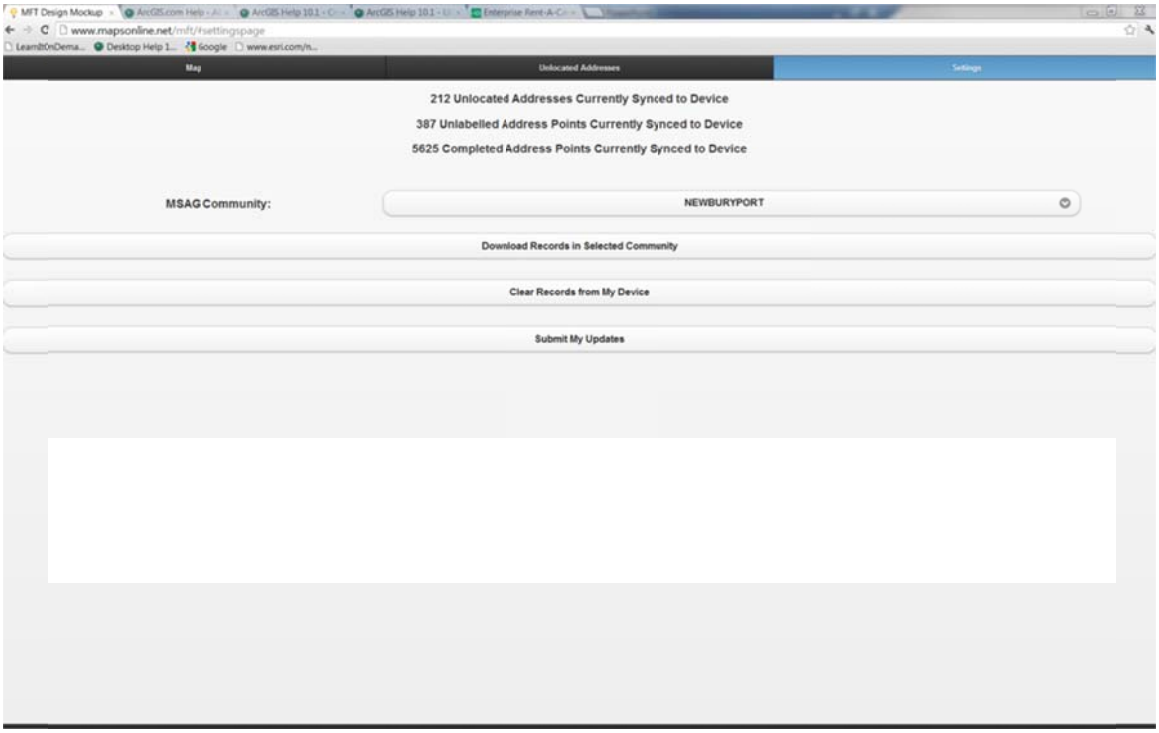
Although this seems like a bewildering variety of use cases, it can be thought of as all possible permutations of basic operations on each of the data sets, points and address records. For points, these are add, delete, explode, select and link selected point to selected address records. For address records, the operations are add, copy and add, edit, delete, select one or more and link selected to selected MP. Linking point location and address record entails selecting a point and one or more address records and transferring the AddrPt_ID from point to address records (see data model on page 13 if this is not clear.)

In addition to editing operations, there are some enhancements to the typical query functions which reflect the grouping of points. When a part of a point is selected, the entire multi-point is selected. In the address list the records linked to a selected point are highlighted and vice versa.

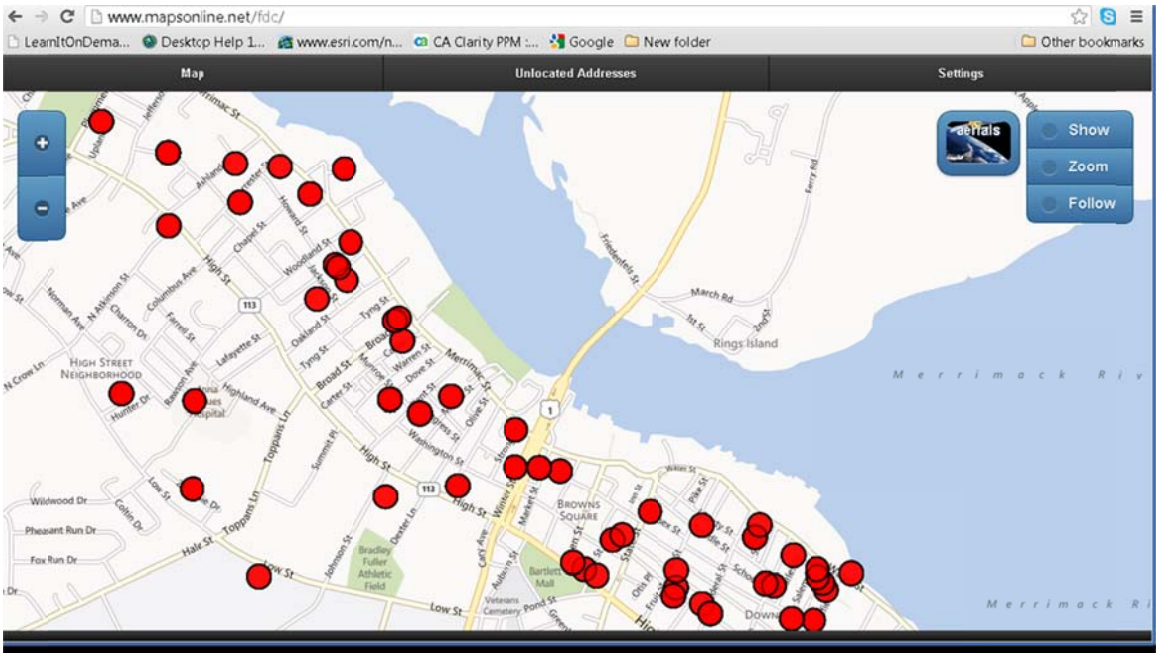
The field data capture application was tested in a browser on both a phone and on a tablet. The initial requirement for the hardware was that it support Android version 4 (Ice Cream Sandwich), since the Chrome browser running on that platform offered the most complete support for HTML5. The browser was developed using existing Javascript libraries to call Open Geospatial Consortium standards-compliant services running at the Commonwealth Data Center in Chelsea MA. A full description of these services can be found at https://wiki.state.ma.us/confluence/display/massgis/Home.

What follows is a series of screen shots showing the use of the field data capture application, as it was deployed in the seven pilot towns. Local officials were accompanied by a regional staff person who was "expert" with the application.

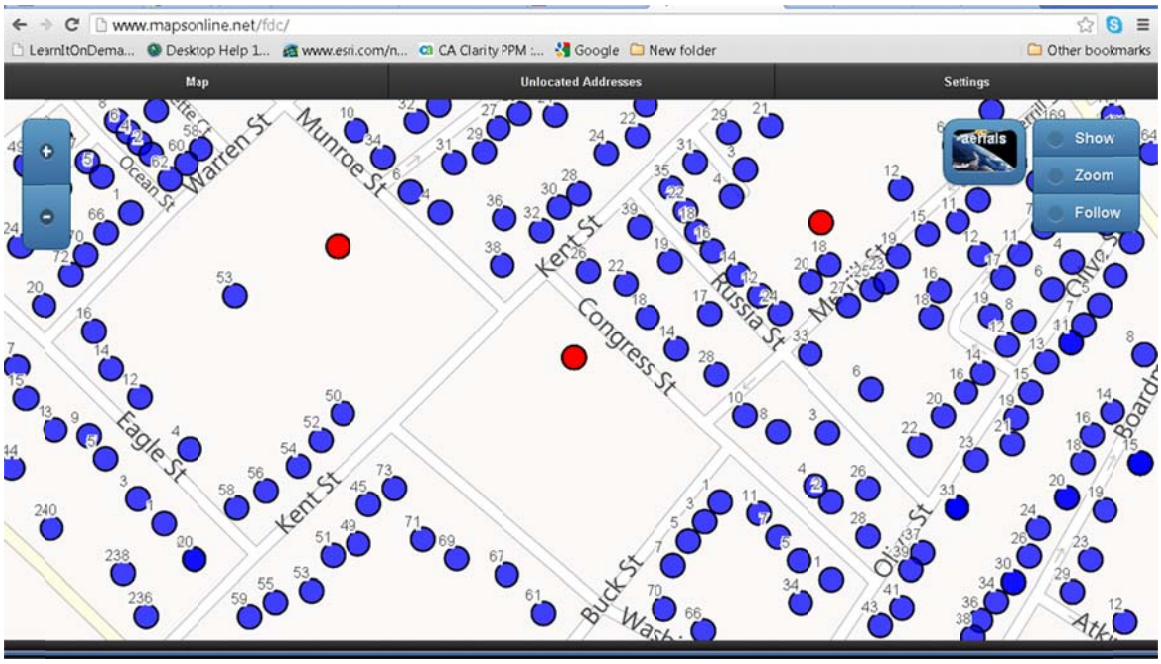
The initial screen offers the user a choice of towns – shown here is “NEWBURYPORT” which was one of the pilot towns.



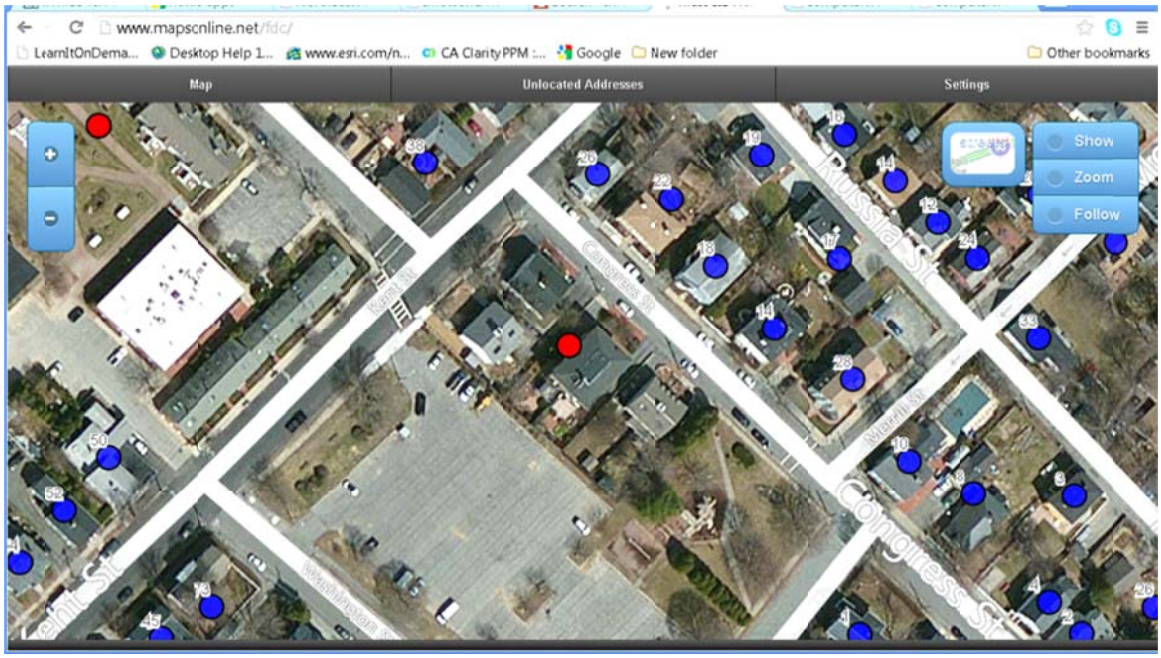
The user then has the option to go to the “MAP” tab or to the “UNLOCATED ADDRESSES” tab. The map tab displays the locations that need to be visited as red dots.



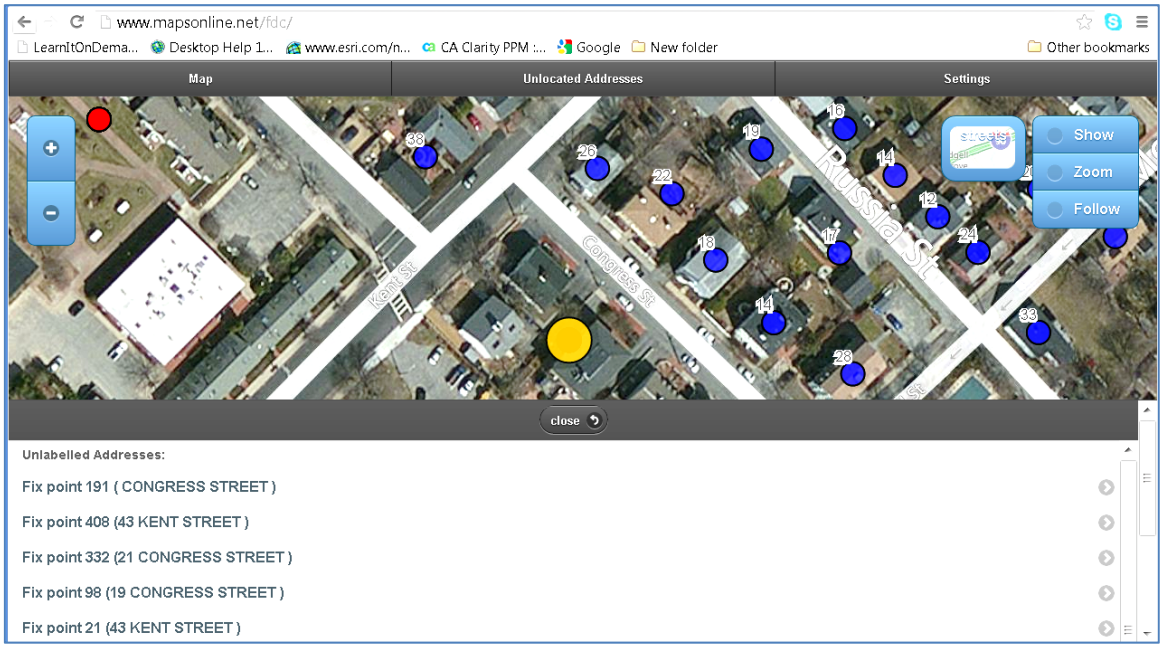
Zoomed in, all the address points appear with street numbers attached.



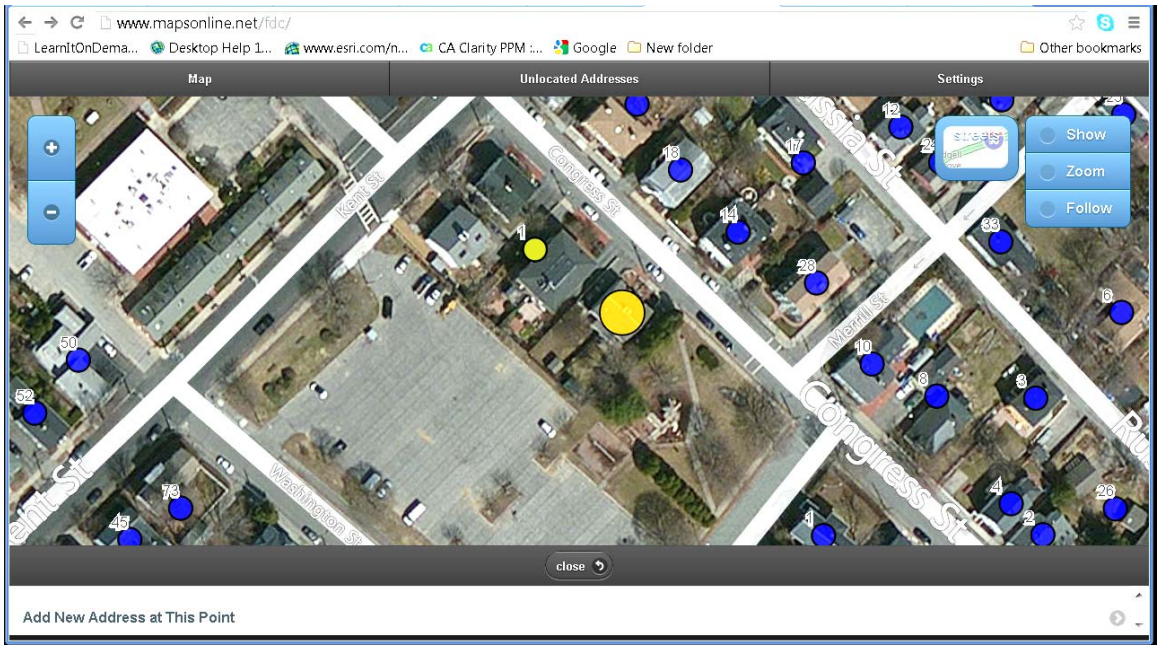
A single location can be viewed on top of the orthophoto – in the image below several structures that need to be differentiated.



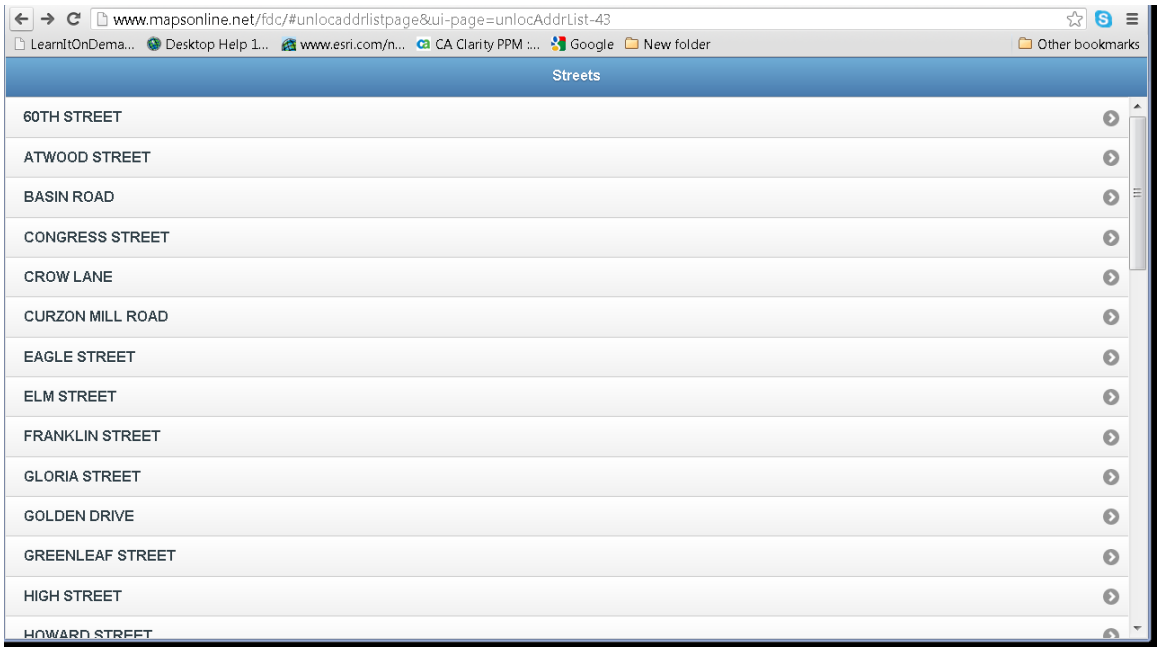
The user has the option to link a point to one of the unlocated addresses as shown below:



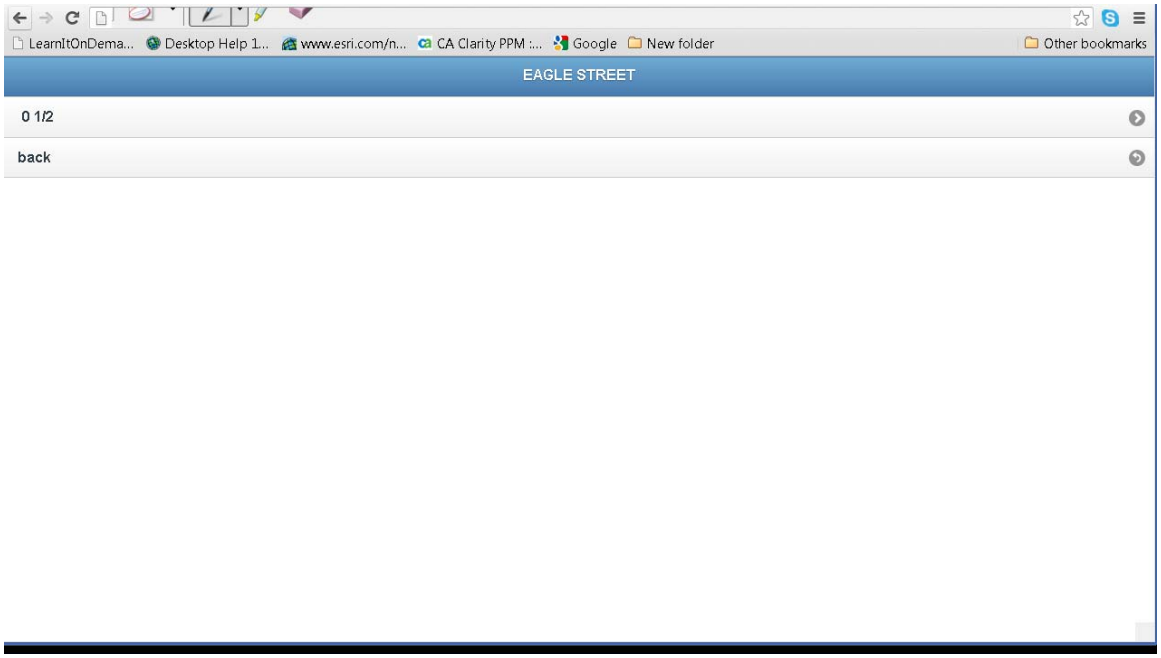
Or to create a new address point:



The address list tab allows the user to pick a street then view address numbers that need to be verified:



These are the numbers on Eagle Street:



Lessons Learned with Mobile App:

There were a number of important lessons learned with the mobile application:

1. **Allow for the unexpected –** we made the mistake of not allowing the user to correct errors that they found independently of the ones that were pre-identified as questions or discrepancies. Mis-spelled or erroneous street names were the best example of this.
2. **Support non-building points –** the full range of address types has to includes such locations as a playground, a parking lot or a utility pole. We did not provide appropriate form fields for entry of these types.
3. **Find the right person to use the app** – given the substantial time commitment, it was worth finding the right person who was not only knowledgeable but also very motivated. We did correctly anticipate that some training and hand-holding were essential. Knowledge and enthusiasm were more important than technical aptitude.
4. **Field work is time intensive** – along the same lines, its important to minimize time in the field by developing as much high-quality data as possible before going out.
5. **Minimize data entry** - always use pick lists but make the number of choices manageable. Related to this is the importance of filtering choices using geolocation – we didn't do this but wish we had.
6. **Give feedback in the user interface** – this is a well-accepted principle of user interface design – not to leave the user wondering whether something has crashed or isn't working right, which often leads them to attempt additional inputs, thus further compounding the problem.