

Investigation of multiple-occupancy addresses

11 May 2004

**Robert James
White Wulf Consultants
Gable Cottage
The Green
Dunham-on-Trent
Newark
Nottinghamshire
NG22 0TU**

email: robertj@cix.co.uk

**Rob Walker
Rob Walker Consultancy
64 Histon Road
Cottenham
Cambridgeshire
CB4 8UD**

email: robwalker@cix.co.uk

Executive Summary

Addresses of multiple occupancy properties, where several addressees share the same address, have been identified in several parts of the Project Acacia work as an area requiring more detailed investigation. Multiple occupancy can be hard to identify and affects an estimated 10-20% of property nationally.

This report is the output of a short study of multiple occupancy addresses in the City of Nottingham. The object of the exercise was to identify and investigate issues associated with these addresses. The report considers some groups of multi-occupancy properties and identifies the different needs and practices of the Acacia partners in recording this type of address in the different source datasets.

A general definition of multiple occupancy has been adopted:

any physical property subdivided by occupation

This encompasses commercial property and flats where there is a lesser level of sharing as well as shared houses and bed-sits.

The address datasets from NLPG, Council Tax, non-domestic rates, OS MasterMap (derived from the Royal Mail PAF) and Land Registry were analysed to find multiple occupancy addresses. It is not straightforward to identify these addresses from these files. Only the PAF and datasets originated from it attempt to identify concealed multiple occupancy (that is one door with many occupants behind it). Some shared multiple occupancy is identified in the VOA data where a premise pays both Council Tax and Non Domestic Rates. In the other datasets, multiple occupancy addresses can only be identified by investigation of the structure and semantics of the address.

The source files are not compatible in structure or content. All have different rules for inclusion of multiple-occupancy addresses. It was found that there are many omissions and duplicates, and business addresses are recorded poorly in all files. All of the files seem to exhibit significant lag between changes and updates.

Only a full comparison of all the files that would involve the standardisation of data structures and re-presentation of data content will enable these issues of quality and consistency to be resolved. An outline of how this might be done is provided. Part of this process is the need to establish standards for data recording that could be applied across all stakeholder datasets. This can be undertaken independently of the wider process.

Much multiple occupancy is not recorded, for example. student halls of residence and care homes of all kinds. Consideration should be given to how those properties in multiple occupation that are of interest to other areas of government could be recorded.

Contents

	Page
1. Introduction.....	1
2. Definition of multiple occupancy.....	2
3. Occurrence of multi-occupancy addresses.....	4
3.1 Identification of multiple occupancy.....	4
3.2 Multiple occupancy flags in the data.....	4
3.3 Multiple occupancy identified by data structure.....	4
3.4 Multiple occupancy identified by semantics.....	4
3.5 Summary of findings.....	5
4. Analysis of multiple occupancy addresses.....	7
4.1 Occurrences in source datasets.....	7
4.2 Examination of multiple-occupancy addresses.....	9
4.3 Semantic v structural analysis.....	9
4.4 Missing Entities.....	10
4.5 Data quality.....	10
4.6 Future activities.....	11
5. Conclusions	12
6. Recommendations	13
Annex A. Definitions of categories of multiple occupation (Residential).....	14
Annex B. Multiple occupancy flags in the data.....	15
Annex C. Multiple occupancy identified by data structure and semantics.....	17
Annex D. Reconciliation of datasets to improve overall quality and interoperability..	24
Annex E. Examples of data relevant to multiple occupancy.....	29

1. Introduction

The Acacia programme is a collaborative project run by a partnership of government agencies, Ordnance Survey, Land Registry, Registers of Scotland, Valuation Office Agency, Improvement and Development Agency (IDeA) and the Royal Mail. It is investigating the requirements for a single national infrastructure for joined-up consistent, high quality, well-maintained addresses and property information.

The project has comprised three main areas of work:

- investigation of specific research topics;
- address data matching trials;
- other investigations.

This work is reported in the Acacia Pilot Project Final Report.

One area that has been identified at several stages as warranting further investigation, has been multiple occupancy properties. These are ones where more than one addressee shares the same basic address. In some cases, these individual addresses are recorded separately, and in others they are all referenced by a single address.

This report is the output of a short study of multiple occupancy addresses in the City of Nottingham. The object of this exercise was to identify and investigate issues associated with such addresses. It considers some groups of multi-occupancy properties and identifies the different needs and practices of the Acacia partners in recording this type of address in the different source datasets.

The source datasets investigated were:

- National Land and Property Gazetteer (NLPG),
- Non-Domestic Rates (NDR),
- OS MasterMap Address Layer,
- Postcode Address File (PAF),
- Council Tax (CTAX),
- Land Register.

The OS MasterMap Address Layer is based upon PAF, which was not used directly in the analyses. In the text, the term OS MasterMap is used when describing the dataset, and PAF when describing the source of the address.

This work has been carried out by Robert James of White Wulf Consultants and Rob Walker of Rob Walker Consultancy over the period March-April 2004.

2. Definition of multiple occupancy

Where multiple occupation is defined in legislation, it is primarily for health and safety and housing purposes (section 345 of the 1985 Housing Act and section 97 of the 2001 Finance Act in particular). It is defined as a single dwelling being occupied by persons that do not form a single household. The English House Condition Survey (EHCS) of 1996 found that there were some 550,000 properties in England in the private rented sector that meet this definition.

This very restrictive definition would exclude business entities and implies a high degree of shared facility within one dwelling. Royal Mail would define multiple occupancy as more than one independent mail recipient behind a single front door and aim to flag all such properties with the number of recipients. VOA files identify where a property (hereditament) pays both council tax and NDR, thus specifically identifying residential/business multiple occupancy. There is no special consideration given in the NLPG or Land Registry files to multiple occupancy, other than the data structures that they use that specifically identify subordinate properties.

Because of these different understandings of multiple occupancy, we have sought out definitions of multiple occupancy that are a little wider and found that one of the best derived from the legislation is used by Nottingham City Council. A copy of this is at Annex A. It contains six categories of residential multiple occupation. We have split Category F into two to recognise that many users would not want to consider a tower block as a single unit but rather as a series of properties. Thus the categories become:

- A.** Bed-sits with shared bathroom and/or kitchen facilities
- B.** Shared houses
- C.** Student and worker accommodation
- D.** Hostels and hotels used for the homeless
- E.** Registered care homes
- F1** flats with third party access to the inside of the property for delivery purposes
- F2** where there is a single point of delivery for all residents.

Additional we need to define a category for business and residential combined multiple occupancy:

- G.** Business premises with residential owners, managers or staff

The word “business” in this context is intended to cover all non-residential property and would encompass hospitals, prisons and properties in the government and voluntary sector.

Purely business properties fall into three categories:

- H.** Shared properties with no particular differentiation (normally associated companies);
- I.** Each business with its own private area, but shared reception and toilet facilities etc;
- J.** Self-contained but with one shared entry area.

These categories are based on the residential categories with H based on B, I based on A, and J based on F. These are first pass definitions and will need to be further considered at a later stage.

The most appropriate definition of multiple occupancy will depend on the particular business purpose for which it is intended. Each of the Acacia partners needs to consider which of the categories they are trying to separately identify, either explicitly or by counts or cross references. There is no reason why any stakeholder should not disregard certain categories or combine them, so long as it is clear what is being done and why.

Taking these eleven categories as a whole, we have created a definition for multiple occupancy:

any physical property subdivided by occupation

In the NLPG, a physical unit would be a master BLPU, and in OS MasterMap data it would be a TOID.

It is clear that subdivision by occupation does not necessarily imply complete physical division or even any physical division of the property. Care should be taken to avoid confusing physical definitions of property with occupational definitions in the same set of data. Although the physically-defined property will often coincide with the occupationally-defined property, this will not always be the case.

3. Occurrence of multiple occupancy addresses in source datasets

3.1 Identification of multiple occupancy

The restricted definition of a house in multiple occupancy found in legislation is intended primarily to cover shared houses, bedsits and even households with lodgers. These are the very categories of multiple occupation that cannot usually be identified from data files that are primarily intended as records of property.

However, in the data, we have identified a range of properties where a single building is subdivided physically in such a way as to have evidence of multiple occupation. There are three ways that a property entry might be identified as being in multiple occupancy these are:

1. flags in the data (PAF/MasterMap and Council Tax/NDR);
2. the data structure;
3. the address semantics.

3.2 Multiple occupancy flags in the data

Two files, OS MasterMap and NDR have multiple occupancy flags. The flag in MasterMap/PAF is used by Royal Mail to identify multiple recipients of mail behind one door. In the NDR case, it is an indicator that identifies that a property occurs both in the Council Tax and NDR files, showing that there is a shared usage between residential and commercial. These are described in Annex B.

3.3 Multiple occupancy identified by data structure

Where the data is provided in a well-structured form such as BS7666, it is quite easy to extract data that matches patterns that are suggestive of multiple occupancy. However with the exception of the NLPG, which is structured in strict BS7666 format, the source data files have much weaker structures. The weakest is the VOA files, where the data as supplied by local authorities has all of the address details for property concatenated into two fields, property name and property number. These are described further in Annex C.

3.4 Multiple occupancy identified by semantics

Rather than identify multiple occupancy from the semantics of the address, it is easier to identify unitary occupancy, or rather what appears to be unitary occupancy, since any property that has no identified subdivisions must be considered as single occupancy unless there is secondary information (flagging) to suggest otherwise. This is described further in Annex C.

There are certain categories of property that can be identified as single occupancy. The two main ones are:

- those identified by a number on the street with no suffixes and no indication of subsidiary premises; and
- those for which there is only a house name.

These categories are easy to identify in all of the files. However what remains is not unequivocally multiple occupancy. It is impossible from the data alone to determine whether “12a” is a subdivision of “12” or is an infill property or for that matter is actually “13”.

3.5 Summary of findings

Whilst the NLPG and Land Registry datasets are readily analysed by data structure, the weaker structure and higher data volatility of the OS MasterMap/PAF and VOA data means that these datasets are best analysed semantically. The files are analysed in detail in annex C. A summary of this analysis is shown in Table 1. This shows the number of singular addresses (those that on the balance of probability are undivided properties), ambiguous addresses (those such as ‘1A’ which might be a flat or an infill property) and the remainder that are assumed to be multiple occupancy.

Table 1. Summary of datasets

	NLPG		VOA		MM		LR	
Singular	122035	82%	115372	90%	118924	92%	121331	91%
Ambiguous	2877	2%	2371	2%	2753	2%	2837	2%
Multiple Occupancy	23978	16%	9736	8%	7398	6%	8859	7%
Total	148890		127479¹		129075		133027	

¹ 5922 addresses in the VOA data and 3001 in the OS-MasterMap data have been excluded because they could not be resolved

One inference that can be drawn from this table, having taken into account the excluded addresses, is that the main difference at this level of resolution is that the NLPG contains 15,000+ more properties than the other datasets, and that these are multiple occupation, as defined by the data structure of the file.

Duplicate addresses have been identified in both the NLPG and the Land Registry data. Subjectively, the level of duplication seems higher in the NLPG. However this alone is unlikely to be the reason all this difference. Some of the additional variation will be because of the inclusion of non-occupiable entities (i.e. not buildings) in the NLPG that have been excluded from NDR in this analysis.

Where flags are used to identify multiple occupancy, there is every reason to believe that these are incomplete, that is they identify only some of the affected properties. In the case of the VOA files, this can be largely corrected by data analysis and the examination of anomalies. However in the case of the Royal Mail originated data (OS MasterMap), this can only be resolved by re-surveying and/or the use of secondary sources.

4. Analysis of multiple occupancy addresses

4.1 Occurrences in source datasets

For there to be a completely meaningful comparison of multiple occupation properties of all kinds between the different files, certain prerequisites need to be met:

- data needs to be in the same data structure (BS7666);
- in all cases of multiple occupation, the same standard of data content should apply whichever system the entry is in;
- internal integrity rules should be applied throughout the data.

Table 2 identifies which multiple-occupancy categories were found in each data file. It was hoped to produce a second table supplying counts of each of these categories, but this task has proved harder than expected and cannot be completed in the time available.

Within the VOA files, certain entities will fall in either Council Tax or NDR because of their treatment for taxation purposes and this is not always dependent on either the status of the occupiers or the physical structure of the building.

The OS MasterMap data flags few of the possible multiple occupied businesses addresses that have been identified as residential/commercial entities in the Council Tax and NDR data. Because of this, we have graded this as B/F in the table to identify that the file aspires to include this data on a flagged basis but only manages in part to do so. Some shared commercial property in all of the files may be identifiable by data analysis that compares counts of primary properties and identifies a multiple occupation where there is more than one incidence of a primary.

Table 2. Occurrence of each type of multi-occupancy address in the source data files.

M-O type		NLPG	OS MM	Land Registry	Council Tax	NDR
A	Bed-sits	B	G	B	B	-
B	Shared houses	G	G	M	M (some)	M (some)
C	Student & worker accommodation	B	G	B	B	-
D/E	Hostels/ Care homes	G	G	G	M (some)	M (some)
F1	Flats with own letter box	M	S	S	S	-
F2	Flats with single delivery point	G	F	B	B	-
G	Business with residential element	B	B/F	B	F	F
H	Shared properties	G	F	G	-	B
I/J	Separate units or shared entry	M	F	B	-	M

Key

G	group entity only
M	mix of group entities and sub-units
S	all sub-units, no group entity
F	group entities flagged as multi-occupancy
B	mix of group entities and some sub-units
B/F	mix of group entities and some sub-units. Some group entities flagged
-	not recorded

4.2 Examination of multiple-occupancy addresses

Some examples of multiple-occupancy addresses are given in Annex E. These examples have been chosen to illustrate some of the difficulties with the data and the differences in styles of recording information between the different files. They cover the following:

- University accommodation,
- Castle Marina – residential boats,
- Sheridan Court – ‘tenement’ numbered blocks of flats,
- Malvern Court – a terrace of houses recorded as if it were a block of flats,
- Magdala Road – houses subdivided into flats,
- Newcastle Chambers - business addresses,
- commercial multi-occupancy on OS MasterMap,
- banks – not specifically a multi-occupancy issue, but a name issue.

The main findings are as follows:

- the datasets have different approaches to recording multiple-occupancy addresses, based upon different business requirements;
- all datasets appear to have missing entries;
- the different ways of recording addresses in each of the datasets means that matching between them can be extremely difficult;
- Where NLPG incorporates the Council Tax data, it carries across errors from that file;
- Not all multiple occupancy is flagged when it should be (OS-MM);
- Business names are not kept up-to-date in the most of the files;
- There is evidence that none of the files copes promptly with property changes, demolition and rebuild or conversion;
- None of the files identify student rooms in halls of residence;
- None of the files copes well with mobile entities (boats in this case).

4.3 Semantic v structural analysis

Two files, NLPG and Land Registry have been analysed structurally and three, OS-MasterMap, Council Tax and NDR have been analysed semantically. It is clear looking at the tables in Annex C, that structural analysis facilitates inter-file comparisons, whilst the different dialects in files prevent semantic comparisons except at the highest levels.

It is also clear when comparing the Land Registry data with the NLPG data, that certain categories although structurally identical must have considerable semantic differences because of the widely different counts for identical categories. An example of this is Building name on street where there is no number and no sub property. The NLPG count is 9667 whilst the Land Registry count is 2575. When the files are examined, the reasons become clearer. NLPG includes descriptive land references for example, as well as non-building tax entities etc.

4.4 Missing Entities

There are a number of entities that do not fit the business needs of data providers but which are of significant importance to government. In some cases these are not recorded at all, in others they are recorded in a manner that satisfies the data providers needs but not the wider needs of the government and commercial communities.

These entities are:

- shared houses and bedsits, (some 550,000+ not recorded at all);
- student halls and similar (based on Nottingham, some 1m + nationally not recorded);
- mobile homes (some 300,000+ nationally recorded in part).

Probably close to 1.75m. households and individuals are not fully identifiable from these data files.

There are many reasons that these need identification:

- electoral registration,
- health and safety,
- fire regulations,
- building regulations,
- national security,
- police matters,
- taxation of all kinds.

Because of the nature of these entities, some could be described in the data e.g. student rooms, and some would need to be flagged with numbers e.g. bedsits and shared houses.

4.5 Data quality

In the course of this analysis, data quality problems have been found in all the datasets.

These include the following:

- omissions
- duplication
- inconsistency of format
- out-of-date names
- lack of currency
- inconsistent use of abbreviations
- concatenation and separation of data fields
- items recorded in wrong fields
- referential integrity failure

It is clear that all the quality issues associated with the source address datasets are especially prevalent in the area of multiple-occupancy. Only a full comparison of all the

files that would involve the standardisation of data structures and re-presentation of data content would enable these issues to be resolved. An outline of how this might be done is at Annex D.

4.6 Future activities

This study has been a preliminary look at the issue of multi-occupancy addresses. It has confirmed the importance of this topic, which has not had a great deal of consideration previously. It is clear that there are many issues, not least due to the different requirements of the various Acacia partners in this area, and the different ways in which they deal with the issues. This includes data maintenance. These issues are not in general restricted to multiple-occupancy, but tend to manifest themselves more strongly in this situation. In the time available, it has only been possible to come up with general conclusions based on one local authority area, and no attempt can be made to quantify the extent of these on a national basis.

5. Conclusions

1. There is no common view of multiple-occupancy among the Acacia partners.
2. All source address files record and differentiate some level of multiple occupancy. Royal Mail and VOA have explicitly identified certain properties. All other multiple occupancy is recorded according to individual business rules.
3. All the datasets store their data in a different format and operate different rules for data content and presentation. Each appears to record different entities, and none records all cases of multiple occupancy in their area of interest.
4. In all the datasets, there is evidence of significant error, omission and duplication.
5. There appear to be a lack of consistent standards for data input, data quality or referential integrity, either in individual files, or for the data providers as a group.
6. The lack of standardisation and data deficiencies leads to significant matching difficulties between systems, uncertainty as to scope of issue (number and types of multiple occupancy) and increased risk of error, mistakes and taxation loss in systems that rely on these data sources.
7. There are questions of currency of the data with many of the parties holding outdated business names, and properties that have been demolished and rebuilt.
8. There is evidence in the Royal Mail data of significant attempts to standardise business names. However, some of the standardisation chosen is trading name and others are legal entity name. In none of the other files is there any evidence at attempting to standardise names.
9. In many of the files, abbreviations are used, particularly in the Council Tax and NDR data, where concatenation has lead to long strings in the name field. Abbreviation hinders interoperability and cross-matching between files particularly where a non-standard abbreviation is used.
10. An estimated one million plus multiple occupancy addresses of interest to government and the wider community cannot be identified from these datasets.
11. All the quality issues associated with the source address datasets are especially prevalent in the area of multiple-occupancy.
12. The Multiple occupancy business data that has been examined suggests that establishing and maintaining a Business file is outside of the current scope of all of these stakeholders and would be a serious task. There seems little doubt that there is a need for such a datasets.

6. Recommendations

6.1 For individual Acacia partners

1. Establish and publish quality control standards for multiple occupancy properties, the basis upon which they collect this data, and the referential integrity rules that apply to their data.
2. Move towards supplying data in BS7666 format even where internal legacy systems hold data in some other format.
3. Undertake a process to improve the quality of their individual datasets and to ensure the highest level of transparent interoperability between data.
4. Aim to collect all the addresses that match their definition of multiple occupancy.

6.2 Collaborative activity

1. The Acacia partners, in conjunction with other major address users, should establish standard definitions of multiple occupancy, so that data can be interchanged between systems.
2. There should be a standard for data content of multiple occupancy property addresses, so that all properties of a particular type would be recorded in the same way across all datasets. This would include the standardisation of the presentation of business names and the removal of all abbreviations.
3. A programme of reconciliation of source datasets should be carried out to improve overall data quality. This should be done using modern enterprise-level software tools rather than on an ad hoc basis.
4. Consideration should be given to the wider needs of government and other sectors particularly but not exclusively in the area of missing properties.

6.3 Further investigations

1. Determine why different datasets that seem to be similar overall, have widely different counts in identical subsets.
2. Better quantify the size of the national problem of multiple-occupancy addresses.
3. Investigate the level of granularity of multiple-occupancy properties that is required by various user organisations and to determine what is sustainable.
4. Investigate the current maintenance cycle and to propose improvements.

Annex A. Definitions of categories of multiple occupation (Residential)

(From Nottingham City Council)

Category A premises

Houses/buildings occupied as individual rooms where there is some exclusive occupation (usually bedroom/living room) and some sharing of amenities (bathroom and or/toilet and/or kitchen). Each occupant lives otherwise independently of each other.

Category B premises

Houses/buildings occupied on a shared basis. These would normally be occupied by members of a defined social group, e.g. students, or a group of young single adults. The occupiers each enjoy exclusive use of a bedroom but would share other facilities including a communal living space.

Category C premises

Houses/buildings with some degree of shared facilities, occupied by people whose occupation is ancillary to their employment or education and is made available through their employer or in connection with a recognised educational establishment.

Category D premises

Houses/buildings generally referred to as ‘hostels’, ‘guest houses’ and ‘bed and breakfast hotels’ or the like. These will provide accommodation for people with no other permanent place of residence as distinct from hotels which provide accommodation for temporary visitors to an area. This category would include establishments used by local authorities to house homeless families pending permanent placement and similar establishments which provide accommodation for people who would otherwise be homeless. It would also include bona fide hotels housing a mixture of homeless households and visitors.

Category E premises

Houses/buildings which require registration under the Registered Homes Act 1984 as amended, providing board and personal care for persons in need, by reason of old age, disability, past or present drug or alcohol dependence or past or present mental disorder.

Category F premises

Houses/buildings which by erection or conversion comprise dwellings which are self contained, all such dwellings comprising accommodation to which access is had via a single entrance door from any common area. Such dwellings would normally contain all the standard amenities but where any might not – and be in an ‘unimproved’ state there would nevertheless be no sharing of amenities with the occupiers of neighbouring dwellings.

Annex B. Multiple occupancy flags in the data

B.1 NDR

The VOA data, NDR and Council Tax, has in its basis the concept of a hereditament, that is, a piece of property which can be inherited. Each hereditament is given a reference number. Where a property is shared between commercial and residential uses, there should be at least one incidence of the reference in each file.

Unfortunately this is not the case. There are 959 entries in the Council Tax data that have references that occur in the NDR file, and 306 of these have negative indicators. In the reverse direction, the 959 matches have 295 negative indicators. Using the flags, 155 flagged records in NDR have no matching records (using the reference) in CTAX and 352 flagged records in CTAX have no matching records in NDR. In many cases these unmatched records can be matched on the address contents but in some they cannot.

Whilst this confusion and failure in referential integrity does not make using this data impossible, it does make it much more difficult and throws up serious questions of quality and reliability.

B.2 OS MasterMap

The multi-occupancy indicator identifies multiple occupiers behind one door. However it is clear that the Royal Mail view is one of recipients and volumes of mail, small and large users, and is not one of households and businesses.

The Royal Mail data identifies only 35 multi-occupancy business properties in Nottingham. These are listed in Table 3. The Largest of these Premier House is a 40 Unit serviced office complex. “officebroker.co.uk” shows 13 more serviced office complexes in Nottingham. Whilst these may be in the 386 multiple occupancy flagged properties without organisation data, the suspicion is that they are simply not identified.

The total Royal Mail multiple-occupancy figure of 421 is less than half the number identified in the NDR file, the majority of which would qualify as being multiple-occupancy in Royal Mail terms. The Royal Mail figure is well below what it should be since the NDR figure excludes any residential-only multiple occupancy which would be expected to be much higher. Whilst indicative, it is not necessarily operationally useful for any purpose where completeness is needed.

From examination of the data, it is clear that there are four main identifiable types of entry flag as multiple-occupancy in the PAF file:

- sub-divided properties identified by number;
- sub-divided named blocks where there is no internal numbering scheme;
- businesses with possible residential sharing;
- business sharing with business.

Table 3. Multiple-occupancy addresses in PAF in Nottingham

FID	NAME	Postcode	Number	comment
1000002308426225	PREMIER HOUSE	NG1 2NA	40	Serviced Offices
1000002307023212	SHOWMANS OVERWINTERING GROUND	NG8 5FH	20	
1000002308426934	DAVID SAIL ,EAST MIDLANDS REGIONAL OFFICE	NG1 6BJ	11	Estate agent
1000002308316000	SPREAD EAGLE	NG7 3PE	9	PUB
1000002308474214	POINTBIO PLC	NG2 4DS	7	
1000002308366777	1 STOP PROMOTIONS	NG1 5BB	6	
1000002308535530	FOSTERS MOBILE WORKSHOP	NG2 4BG	6	
1000002308429913	NOTTINGHAM CITY COUNCIL	NG1 4BT	6	Government
1000002308365959	MAIL BOXES ETC	NG1 5FQ	5	
1000002308919168	THE BUSINESS CENTRE	NG1 1LH	4	College
1000002308534433	PARK YACHT INN	NG2 4DS	4	PUB
1000002308433832	DENTAL SURGERY	NG3 5AA	4	Dentist
1000002308433862	HIMALAYA TANDOORI RESTAURANT	NG1 3FT	4	Restaurant
1000002308362493	DENTAL SURGERY	NG1 5DN	3	Dentist
1000002307023361	BEE ENTERPRISES WORKSHOP	NG8 6GD	3	
1000002308427084	BENEFITS AGENCY	NG2 3BZ	3	Government
1000002308476115	TERRYS BUTCHERS	NG2 4LW	3	Shop
1000002308430377	BENEFITS AGENCY	NG1 6AX	3	Government
1000002308316310	IKANO FINANCIAL SERVICES LTD	NG7 5LN	2	
1000002308312143	DENTAL SURGERY	NG7 1PZ	2	Dentist
1000002308162070	DIRECT VALUATIONS	NG9 2NX	2	
1000002308362936	EUROPADISC	NG1 6JE	2	
1000002308323648	VISTA SIGNS	NG7 7DA	2	
1000002308324009	MARK ONEDENTAL SURGERY	NG7 7DA	2	Dentist
1000002308935929	HORSE AND GROOM	NG7 7EA	2	PUB
1000002308480928	M M CLOTHING	NG3 1JE	2	
1000002308476111	CORAL	NG2 4LW	2	Shop
1000002308476112	NAIKS NEWSAGENT	NG2 4LW	2	Shop
1000002308536222	TIGER EUROPEAN	NG2 4AP	2	
1000002308536295	JOHNSONS BAKERS	NG2 4DE	2	Shop
1000002308536236	MAID MARIAN STORES	NG2 4AX	2	Shop
1000002308939354	LIGHT YEARS LTD	NG2 4AX	2	
1000002308423146	P T STORES	NG2 2LH	2	Shop
1000002308430375	INDEPENDENT TRIBUNAL SERVICE	NG1 6AZ	2	
1000002308438396	EL SOMBRERO	NG5 2FS	2	Restaurant

However except where a secondary source has been used, outside of stakeholder data sets, each of these divisions is supposition from the address semantics and the reality may in fact be different from what appears in the address.

Annex C. Multiple occupancy identified by data structure and semantics

C.1 NLPG

Since the BS7666 structure of the NLPG separates Secondary Addressable Objects (SAO) from Primary Addressable Objects (PAO). All those objects that are correctly recorded as SAOs are properties in multiple occupation. There are two main reasons why a non-multiple occupied property might be wrongly recorded as a SAO, error in creation and complex addresses that does not fit into BS7666. These should be identifiable using semantic analysis of data content.

Most error occurs where the semantics of an address are ambiguous, for example Courts can be Streets and Blocks and if a ‘street’ court is not identified as such it may lead to SAOs being created. Errors can occur the other way round where a SAO has been numbered on a street and the block that it falls in is not identified. The Malvern Court example in the annex F shows properties that are identified as flats (SAOs) but where each has a different street number on Mapperley Road. In reality, these are large townhouses and are in no sense SAOs of Malvern Court. Malvern Court is in effect a Terrace on Mapperley Road and as such is difficult to fit into a BS7666 structured file.

In some cases address complexity will lead to there being insufficient space in the BS7666 structure to record a property without putting a PAO in the SAO field. A typical example of this would be a small block of cottages on a road where the address was say:

2 Rose Cottages, Laneham Road

where there are a number of choices as to how to record the address. The most frequently used would be to record the “2” as the SAON and Rose Cottages as the PAON. This clearly would record a PAO as an SAO and imply multiple occupancy.

Table 4 shows a breakdown of the NLPG data based on the structure of the record. This is the NLPG without any alternate and historic records, which have been resolved. Because of inconsistencies in the allocation of BLPU and LPI status codes, certain BLPUs are not resolvable, that is they have more than one equally valid candidate for being the “correct address”. These records have been excluded.

A number of abbreviations have been used to identify the address structure these are

N	– Street Number
Range	– number range on street
BNAM	- Building Name
N+A	- Number with alpha suffix
SNUM	- sub-building number
SNAM	- sub-building name

This is a contraction of what would be expected in BS7666, particularly in respect of sub-building ranges and suffixes which are present in such few numbers that identifying them adds little to the analysis.

Table 4. Analysis of addresses in NLPG

Address Structure	Number	%
N	110468	74%
RANGE	1192	1%
RANGE,BNAM	125	0%
BNAM	9667	6%
N,BNAM	583	0%
Total Singular	122035	82%
N+A	2877	2%
Total ambiguous	2877	2%
BNAM,SNAM	5305	2%
BNAM,SNAM,SNUM	182	0%
BNAM,SNUM	9458	6%
N,,SNAM,SNUM	0	0%
N,BN,SN	202	0%
N,BN,SNAM	696	0%
N,BN,SNAM,SNUM	1	0%
N,SNAM	7104	5%
N,SNUM	10	0%
RANGE,BNAM,SNAM	212	0%
RANGE,BNAM,SNAM,SNUM	0	0%
RANGE,BNAM,SNUM	7	0%
RANGE,SNAME	801	1%
RANGE,SNAME,SNUM	0	0%
Total multiple Occupancy	23978	16%

Certain of the possible combinations have been left in the table even though the numbers identified are zero, to facilitate comparison with the Land Registry file that is analysed in the same way.

C.2 Land Registry

The data structure of the Land Registry file contains elements of BS7666 structure and Royal Mail structure. There are two street fields but these are handled in a different manner to the PAF data in that where a secondary street is present, this is recorded in the street field and the main street is “promoted” to the second street field. This is the opposite of the PAF and adds to the complexity of handling this data since both fields have to be searched to find all the properties relating to one main street.

The primary address elements are recorded to the detail of BS7666 whilst the secondary address detail is split only into name and number.

Like BS7666 and unlike Royal Mail, there is no specific field(s) for the identification of organisations/businesses. However a significant number of properties are identified by company name alone in the Name field, more than 500 (around 5%) of all business premises (based on NDR).

In general there seems to be less data volatility in the Land Registry files at a property level. What volatility there is seems to be small in volume and essentially error.

Table 5 is an analysis of the Land Registry data on the basis of data structure

Table 5. Analysis of addresses in Land Registry

Address Structure	Number	%
N	117573	88%
RANGE	850	1%
RANGE,BNAM	30	0%
BNAM	2575	2%
N,BNAM	303	0%
Total Singular	121331	91%
N+A	2837	2%
Total ambiguous	2837	2%
BNAM,SNAM	194	0%
BNAM,SNAM,SNUM	2693	2%
BNAM,SNUM	1358	1%
N,SNAM,SNUM	3052	2%
N,BN,SN	60	0%
N,BN,SNAM	27	0%
N,BN,SNAM,SNUM	615	0%
N,SNAM	522	0%
N,SNUM	47	0%
RANGE,BNAM,SNAM	8	0%
RANGE,BNAM,SNUM	1	0%
RANGE,BNAM,SNAM,SNUM	81	0%
RANGE,SNAME	35	0%
RANGE,SNAME,SNUM	166	0%
Total multiple Occupancy	8859	7%

The abbreviations used are the same as the NLPG.

C.3 OS MasterMap/PAF

The data in the address layer of the OS MasterMap product follows the structure used by the Royal Mail. This is similar to BS7666 except that there is a secondary thoroughfare field, and the primary and secondary addressable objects are split into only number and name. This means that number ranges and alpha suffixed numbers have to be recorded as names. Additionally there are fields for organisation and department names.

In theory this should mean that the data in OS MasterMap can be converted into a BS7666 format relatively easily. However, this is not the case because even within the same small area of Nottingham, identically formatted addresses may be recorded in different ways. In the data, two examples of this are industrial estates that are recorded both as secondary thoroughfares and as building names, and blocks of cottages that are recorded as building names with preceding number in the same field, as primary and as secondary thoroughfares.

Whilst these variations do not prevent conversion to BS7666 format, they do make the task considerably more difficult. It should be emphasised however that since the Royal Mail data is primarily intended for mailing, none of these formats are in any sense unsuitable for that. However this variation in recording is less acceptable if the file is to be used for analysis or cartographic purposes.

Because of these difficulties, OS MasterMap data was analysed on the basis of the semantics of entries, although often the effects of the semantics is altered by the data structure. For example, "George Court" in the street field would be accepted as a street, whilst in the Building name field it would be a building, and if it fell into the secondary thoroughfare field it would be ambiguous though would have to be treated as a street. Results are shown in Table 6.

The categories that emerged from the analysis fell into five groups:

- properties that are on the balance of probability, singularly occupied;
- properties that are ambiguous and can only be resolved by secondary information;
- property groups that are ambiguous and can be resolved by further processing;
- properties that are in multiple occupation;
- residue of unresolved entries.

The result is that in the OS MasterMap file, and by inference the PAF, some 6% of properties can be identified as multiple occupancy and that the majority of these are identified by having the word 'Flat' or 'Apartment' either on their own or associated with a number or a single letter, and where there is an identifiable named or numbered property in which the flat or apartment falls.

Table 6. Analysis of addresses in OS MasterMap

Type of address	No.	%	Note
Number on street	113896	86%	
Named building with and without number	503	0%	
Organisation name no other details	1400	1%	
Organisation name and building number	2775	2%	
Units in name	278	0%	
Total Singular	118924	90%	
Number with alpha suffix	2535	2%	
Alpha suffixed numbers with Organisation	218	0%	
Named building with street number mostly Organisation	332	0%	*
Number ranges mostly with subsidiary elements and organisation names	938	1%	*
Number and name concatenated in name field	585	0%	*
Organisation in named building	604	0%	*
Total ambiguous	5212	4%	
Number in block	914	1%	
Units in building	352	0%	
Flat and apartments	5782	4%	
Numbered properties flagged as multiple occupancy	350	0%	
Total multiple occupancy	7398	6%	
Residue with organisations	346	0%	
Residue	196	0%	
Total Residue	542	0%	

* these ambiguous sets can be refined by data processing that is if the same primary property occurs more than once it must be multiple occupancy

C.4 VOA- Council Tax and NDR

These two files have the same structure and the data has been manipulated in similar ways to fit this structure. There is a 'Firm' field and three "street" fields, but primary and secondary address objects have been forced into two fields. An issue with this split into number and name is that it seems that incoming data is concatenated and then any number at the start of the concatenation is put into the number field even where this is inappropriate. For example, "2nd floor flat" would be recorded as "2" in the number field and "nd floor flat" in the name field. Whilst this is acceptable for creating mailing labels it is not acceptable for other purposes.

An extreme case of the effects of concatenation is:

TUTORS FLAT GND FLR STAIRCASE XII LENTON AND WORTLEY HALL

where four elements have been concatenated into the name field. Whilst it is not impossible to re-parse this into a reasonable structure, it would be both difficult and risky.

Similarly to the Royal Mail file, entities float between fields so streets may appear in any of the three street fields and in the name field. Buildings can occur in the secondary and tertiary street fields. This does not necessarily break the systems internal rules, since it seems that the secondary “street” field is intended to hold block names, the tertiary “Street” field is little used.

The effect of this volatility of data placement to fields is to make conversion into a standard format for comparison with other datasets difficult.

For this reason, these two files have been analysed semantically as can be seen in the tables 7 and 8, although some of the categories are similar to those in the structured analysis of NLPG and Land Registry, it is clear that when comparing the OS MasterMap and these files that each file has a different dialect that makes comparison difficult.

Table 7. Analysis of addresses in Council Tax

Type of address	No.	%	Note
Number on street	109282	89%	
Named building with and without number	607	0%	
Total Singular	109889	89%	
Number with alpha suffix	2095	2%	
Total ambiguous	2095	2%	
Number in block	2185	2%	
Flat and apartments	6629	5%	
boats	9	0%	
Flats identified by the word Floor with 2 nd , 3 rd etc	346	0%	
Total multiple occupancy	9169	7%	
Blank records	696	1%	
Residue	1147	1%	
Total Residue	1843	2%	

Multiple occupancy figures in this file are somewhat higher than the OS MasterMap data, although broadly in line with the Land Registry data. Both of these are substantially lower than the NLPG figures. Additionally, the file contained 696 records that were blank for all of the address fields

Table 8. Analysis of addresses in NDR

Type of address	No.	%	Note
Number on street	2905	28%	
Named building with and without number	1510	15%	
Company name on street	400	4%	
ranges	668	6%	
Total Singular	5483	53%	
ranges with company name	40	0%	
Number with alpha suffix	236	2%	
Total ambiguous	276	3%	
units identified by the word Floor with 2 nd ,3 rd etc	540	5%	
Total multiple occupancy	540	5%	
adverts	293	3%	
Residue	3786	36%	
Total Residue	4079	39%	

The NDR file is by far the hardest to work with because of the extreme concatenation of items that could be in five or six fields into two.

For this reason the results of this table should be treated cautiously because of the very high level of residue, unresolved addresses, and in particular the fact that the named building category contains a lot of compound name and business entries.

Annex D. Reconciliation of datasets to improve overall quality and interoperability

The issue of handling multiple data sets and merging them on output whilst holding the originating data in unchangeable legacy systems, is not a new one. A number of companies produce middle-ware products to undertake this task. These products, usually referred to as ETL tools or packages, are Extract, Transform, Merge, Load and Manage software. Essentially these packages manage metadata about each file on the basis of not just what should be but what is. They allow the data files to be transformed both in structure and content following sets of user-defined rules which are often global but in some cases will specify the transformation of a single difficult record.

Thus a series of files with differently formatted data and different creation rules can be standardised and new output files can be created that meet a series of coherent rules. These can then be de-duplicated and merged to create a new data file. The five largest suppliers in this area by market share are Informatica, SAS Institute, Ascential Software, Evolutionary Technologies International and IBM.

The flow chart in Figure D illustrates how such a product could be used in with other tools and techniques to create and maintain a National Address Dataset.

These processes have been presented in such a way that any single file owner could follow processes 1-4 independently of all other files but processes 5 and 6 require that two or more files are available for comparison. The case that the more files that are used in this process, the greater chance that the file will be complete, and also the greater the probability of identifying errors and mistakes in the files.

If the aim is to produce standard national data for the national benefit, it is clear that the activities would have to be controlled by some agency of government. Whilst the actual processing could be out-sourced there needs to be clear and unequivocal ownership of the procedures within government.

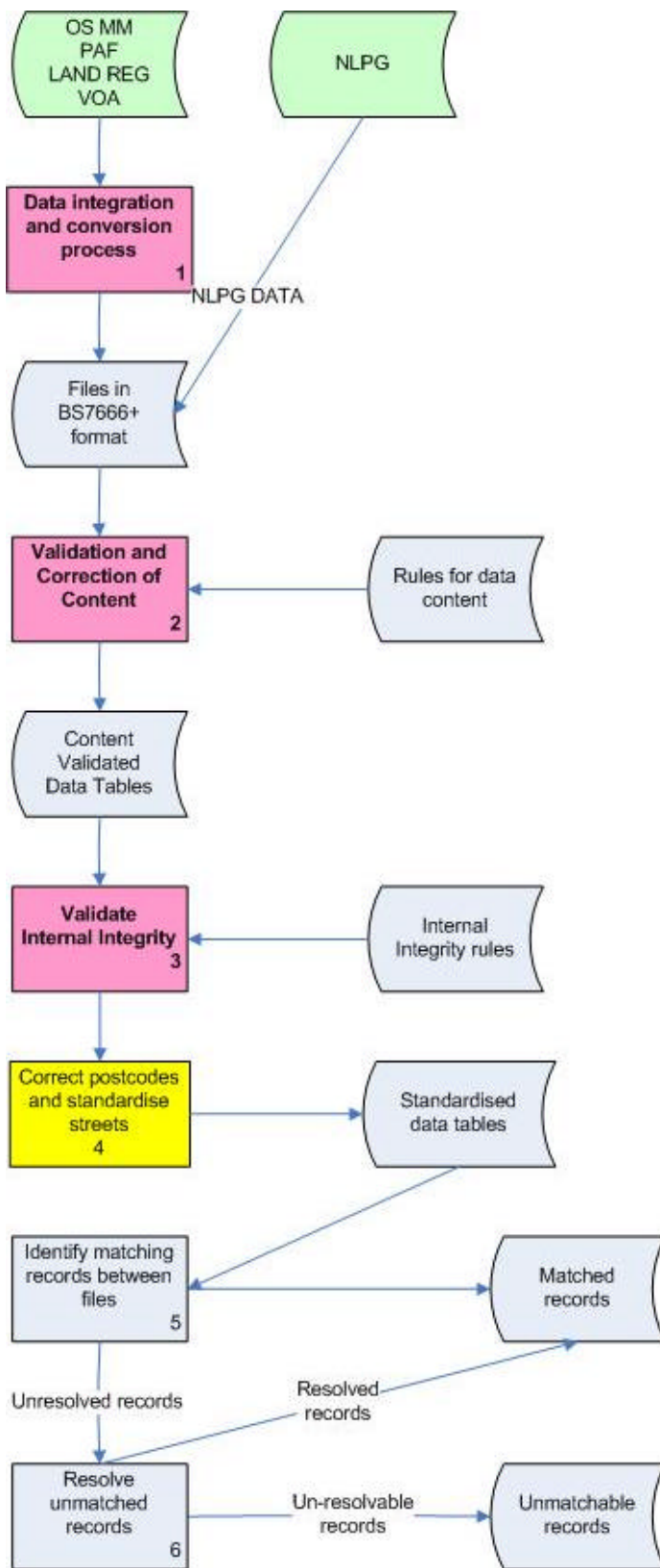


Figure D. Process for creation of combined address database

Process 1 – Data integration and conversion

This process involves the reformatting of data so that it meets or exceeds the standards laid out in BS7666. This will involve the parsing of most addresses and the correction of some of the presentational errors that would prevent accurate parsing. The output from this process will be a file in BS7666 format. The NLPG would not have to go through this stage since it is already in BS7666 format.

Process 2 – Validation and correction of content

This research study has shown that there are problems with how data items are presented this issue primarily effects multiple occupancy and business addresses where there is the greatest scope for different presentation. A prerequisite of this stage is that definitive rules are set covering all aspects of address data presentation within the BS7666 structure. Some issues that will have to be covered include:

- spelling,
- punctuation,
- abbreviation,
- standard presentation of Company names,
- use of individuals names in data,
- representation of rooms in larger units.

These rules will have to be tightly written as they will have to be operated within an unforgiving computer environment.

Process 3 – Validation of internal integrity

All the datasets have a series of implicit rules. There is little evidence that these are checked rigorously in any of the systems. These internal integrity rules have to be extracted from the data and will have to address amongst others the following issues:

- Referential Integrity (where one data element relies on another any change in the later must be reflected in the former);
- Range integrity (a range should end higher than it starts and should not include numbers identified elsewhere individually);
- Elemental integrity (where an element is authoritatively defined e.g. the postcode, it should be checked against the relevant definitive file);
- Group Integrity (where a group of properties share common features these common elements should be identical throughout the group).

Process 4 – Correct Postcodes and standardise streets

In each of the files there are several hundred postcodes that do not exist in OS MasterMap Royal Mail derived data and often thousands of records that have no postcode, 5984 in the NLPG. This is a weakness that can easily be corrected using standard Postcode cross-matching software available from many Royal Mail authorised value added resellers of the PAF.

One of the benefits of having accurate validated postcodes on the data files is that since each postcode can only apply to one street, accurate postcodes allow the additional of postal standard street, locality and town names.

There is no evidence that these are any better than alternatives in other files. However, the existence of software to achieve this standardisation and the fact that in most cases the primary use of this data will be for mailing purposes, suggests that this is a good initial approach to standardising the separate files.

Process 5 – Identify matching records between files

This is a process that would have to be driven by the custodian of these processes using data warehousing/database management software possibly with the use of various software tools. The aim is to identify only those addresses that unequivocally matched between files.

In this process, one file would have to be chosen as the base file and having rigorously de-duplicated this file, the other data files would be matched against it. At the initial stages, simply matching against the initial base file would be sufficient to declare a match since unmatched addresses when they are resolved would highlight any erroneous addresses in the base file. One major problem with these datasets on a national basis is that many of the addresses in all the databases originated in the PAF, which being sourced originally from local authority data leads to the propagation of identical errors across the databases. Thus correctness cannot be decided by a simple vote of the datasets.

Process 6 – Resolve unmatched addresses

Addresses that do not match to the base file will be of three types, addresses that:

- a) exist in one file but do not exist in all files for example ATM's;
- b) should be in the base file but are missing;
- c) are in the base file but errors in one or both of the files prevent matching.

The task of this process is to resolve these un-matched addresses. Type "a" unmatched addresses would be added to the file of unmatchable addresses. Type "b" addresses would be added to the base file. Type "c" addresses would be resolved if at all possibly without referring to the originator for resolution and if this cannot be achieved they would be referred to the originator.

Unmatchable data is not made up of failed records, since the assumption is that these be held in suspense files and would be passed back to their originators for correction, they are rather records that exist in one file but should not exist in some or all of the others. Examples would include taxation entities that occur in the NDR file such as advertising rights or PO boxes in PAF originated data.

Additional Data files

There would be benefits in bringing files together other than simply those that formed the basis for this research. The Electoral Roll, files held by utilities and credit reference

agencies etc. would all enhance the quality and completeness of this approach, however many if not all of these people-related files may not be available to the Acacia partners.

The concept of the process as a whole is that by standardisation we achieve internal consistency and quality management within each of the files and open up the possibility of cross comparing the files.

Outputs

The outputs from this process would be:

- standardised copies of each participating data file;
- tables of cross-matching references between records in files;
- a table of unmatchable data;
- reports of errors would be passed to participants;
- national file of proved and consistent addresses.

Maintenance

One of the benefits of ETL tools is that they are generally designed to both be used for a one-off conversion and to act as a long term bridge between legacy and new systems. As such they would allow the benefits of all the datasets to be made available particularly to Government without requiring any changes in existing systems.

It would be necessary however to keep these standardised files and the new national address file, and all the other data in such a way that it could readily be supplied to users move in government and beyond, this would require an enterprise-level server and software. The software would need to encompass both database management and data warehousing and would have to be completed with data mining/OLAP tools to be used by the managers of the processes to ensure quality and consistency.

Annex E. Examples of data relevant to multiple occupancy

E.1 University Accommodation

The University of Nottingham halls of residence are listed in Table E.1. This shows the Council Tax band, and the number of entries in each source dataset, OS MasterMap, Council Tax, Land Registry and NLPG, and the number of rooms in the Hall (obtained from the University web site).

Table E.1. University of Nottingham Halls of Residence

	C Tax Band	H	MM	CTax	LR	NLPG	No of rooms
ANCASTER HALL	H	✓	1	9	1	10	272
BROADGATE PARK			1	2	1	1	1377
CAVENDISH HALL	H		1	6	1	6	279
CRIPPS HALL	H	✓	1	17	1	28	301
DERBY HALL	H	✓	1	11	1	11	324
FLORENCE BOOT HALL	H	✓	1	9	1	9	194
HUGH STEWART HALL	H	✓	1	9	1	9	347
LENTON AND WORTLEY HALL	H		0	5	0	4	303
LINCOLN HALL	H	✓	1	10	1	10	221
MELTON HALL	G		0	1	0	1	0
NEWARK HALL	G		0	11	0	10	400
NIGHTINGALE HALL	H		1	4	1	4	148
RUTLAND HALL	H		1	9	1	9	271
SHERWOOD HALL	H	✓	1	10	1	9	267
SOUTHWELL HALL	G		0	6	0	6	193
WILLOUGHBY HALL		✓	1	8	1	10	260
			13	127	12	137	5377

This example was chosen to identify the fact that none of the files record the detail of student accommodation. The largest University of Nottingham student unit Broadgate Park is shown in Figure E.1 just north of Broadgate the B6464. There are 1350-plus students in this unit alone. Broadgate Park has many separate buildings and a number of internal roads.

Differentiation down to block and room level would be required for a number of purposes. In the Electoral Roll, this level of fineness of recording would be needed to reduce scope for impersonation and other electoral fraud.

The student accommodation at Nottingham University alone represents 1.5-2% of the total population of Nottingham. If other similarly under-recorded privately owned student accommodation and halls at Nottingham Trent University are added, the figure will be at least twice this.



Figure E.1 University of Nottingham campus.

(Map based upon the 1991 Ordnance Survey Landranger 1: 10 000 scale plans with the permission of Ordnance Survey on behalf of the Controller of Her Majesty's Stationery Office, Crown Copyright. University of Nottingham, Licence number ED100017891)

E.2 Castle Marina

The castle marina example (Figure E.2) is shown as an example of the recording of mobile entities that would include residential boats of all types as well as caravans and trailers. It is difficult to find figures for the number of people who live in mobile homes. However one of the Park Homes Owners associations estimate a figure of 200,000 trailer homes. When less permanent caravan dwellers and boat dwellers are added to this figure, we are probably looking at 300,000+ homes or around 1% of the national stock.

The data for castle Marine shows some 25 residential boats out of a marina that from the associated aerial photograph appears to have in excess of 100 moorings. The remainder of the moorings may be for boats that are used only for recreational purposes. It is impossible simply by looking at a boat to determine which of these cases is true.

Table E.2 is based on entries in the NLPG plus two boats identified in the council tax file, Parfie and African Queen that are not included in the NLPG and so do not have UPRNs.



Figure E.2 Castle Marina
(provided by Nottingham Castle Marina)

The four columns on the right of the table identify the presence of these boats in other files. It is clear that with the exception of Council Tax and NLPG, recording of these entities is restricted to the highest level of marina only.

Table E.2 shows also that many of the entries are duplicated in the NLPG with different UPRN's and two names are used for the street that the marina is on. In some cases, as well as the boat name, a berth number is given. This is typical of a mobile home problem in that the "home" has an identity of its own and that coupled with a berth plot number, identifies the address. In many cases only one of these is available and this may make interoperability between systems and de-duplication within a system problematic.

To identify the duplicated records and ones with missing elements, a considerable number of searches through the data had to be made and the data had to be tidied up to enable these issues that are not apparent from a simply scanning the data to be identified.

Table E.2 Castle Marina – A Mobile Home Example

UPRN	BERTH	NAME	PAON		OS-MM	Land Registry	Council Tax	NDR
100031530075	E13	ALMOST	CASTLE MARINA	CASTLE BOULEVARD			X	
200001410740	E13	ALMOST	CASTLE MARINA	CASTLE BOULEVARD			X	
200001407044		BOHEMIAN	CASTLE MARINA	CASTLE MARINA ROAD				
100031530083		BRAMBLE	CASTLE MARINA	CASTLE BOULEVARD				
200001407037		BRAMBLE	CASTLE MARINA	CASTLE MARINA ROAD				
100031530082		CASTLE EXPLORER	CASTLE MARINA	CASTLE BOULEVARD				
200001407043		CASTLE EXPLORER	CASTLE MARINA	CASTLE MARINA ROAD				
200001410752		CASTLE EXPLORER	CASTLE MARINA	CASTLE BOULEVARD			X	
100031530074	G8	DISCOVERY	CASTLE MARINA	CASTLE BOULEVARD			X	
100031530080	C35	DREAMBOAT	CASTLE MARINA	CASTLE BOULEVARD			X	
200001410741	C35	DREAMTIME	CASTLE MARINA	CASTLE BOULEVARD				
200001410747	G14	DUB TUB	CASTLE MARINA	CASTLE BOULEVARD			X	
100031530071		DUB TUB	BOAT DUB TUB CASTLE MARINA	CASTLE BOULEVARD			X	
100031530084		HARDY	CASTLE MARINA	CASTLE BOULEVARD			X	
100031530077	C16	HILDEGARD	CASTLE MARINA	CASTLE BOULEVARD			X	
200001410751		KIWI VIKING	CASTLE MARINA	CASTLE BOULEVARD				
100031530085		KIWI VIKING	CASTLE MARINA	CASTLE BOULEVARD				
100031530073		LA SHAK	LA SHAK TOW-PATH MOORINGS	CASTLE BOULEVARD			X	
200001410749		LA SHAK	TOW PATH MOORINGS	CASTLE BOULEVARD			X	
200001410745		LESLEY ANN 3	CASTLE MARINA	CASTLE BOULEVARD			X	
100031530081		MARGARETTA	CASTLE MARINA	CASTLE BOULEVARD				
200001407042		MARGARETTA	CASTLE MARINA	CASTLE MARINA ROAD				
200001407046		MERLIN	CASTLE MARINA	CASTLE MARINA ROAD				
100031530078	C21	MISTY	CASTLE MARINA	CASTLE BOULEVARD			X	
200001407038	C21	MISTY	CASTLE MARINA	CASTLE MARINA ROAD			X	
200001410744	C21	MISTY	CASTLE MARINA	CASTLE BOULEVARD			X	
100031530086		NICE AND BRIGHT	CASTLE MARINA	CASTLE BOULEVARD			X	
200001407041		NICE AND BRIGHT	CASTLE MARINA	CASTLE MARINA ROAD			X	

200001410750		NICE AND BRIGHT		CASTLE MARINA ROAD			X	
100031530079	C24	NICOLA II	CASTLE MARINA	CASTLE BOULEVARD			X	
200001410742	C24	NICOLA II	CASTLE MARINA	CASTLE BOULEVARD			X	
100031530087		PATRICIA M TOO	CASTLE MARINA	CASTLE BOULEVARD			X	
200001407045		PATRICIA M TOO	CASTLE MARINA	CASTLE MARINA ROAD			X	
200001410743	E14	PORTLAND BILL	CASTLE MARINA	CASTLE BOULEVARD				
100031530076	E14	PORTLAND BILL	CASTLE MARINA	CASTLE BOULEVARD				
200001410746		ROSE	CASTLE MARINA	CASTLE BOULEVARD				
100031530072		ROSE CASTLE	BOAT ROSE CASTLE MARINA	CASTLE BOULEVARD				
200001407040		TIGGER II	CASTLE MARINA	CASTLE MARINA ROAD			X	
200001410748	C30-32	TIGGER II	CASTLE MARINA	CASTLE BOULEVARD			X	
200001411777		TOW PATH MOORINGS	TOW PATH MOORINGS	CASTLE BOULEVARD				
200001407039		WATERMINT	CASTLE MARINA	CASTLE MARINA ROAD				
		PARFIE	CASTLE MARINA				X	
		AFRICAN QUEEN	CASTLE MARINA				X	
100032287943			CASTLE MARINA PARK	CASTLE BOULEVARD	X	X	X	X
100032309126			CASTLE MARINA	CASTLE BOULEVARD	X	X	X	X
200001412662			CASTLE MARINA	CASTLE MARINA ROAD	X	X	X	X

Table E.3 Sheridan Court – an example of tenement style numbering (Floor+Flat)

LR Sub No	Sub name	CTax no	Name	OS-MM	NPLG Group 1 UPRN	SAON	NPLG Group 2 UPRN	SOAN
101	FLAT	1	-01 SHERIDAN COURT 66	FLAT 101	100032093763	FLAT 101	200001408392	1 1
102	FLAT	1	-02 SHERIDAN COURT 66	FLAT 102	100032093754	FLAT 102	200001408393	1 2
103	FLAT	1	-03 SHERIDAN COURT 66	FLAT 103	100032093755	FLAT 103	200001408394	1 3
104	FLAT	1	-04 SHERIDAN COURT 66	FLAT 104	100032093756	FLAT 104	200001408395	1 4
105	FLAT	1	-05 SHERIDAN COURT 66	FLAT 105	100032093757	FLAT 105	200001408396	1 5
106	FLAT	1	-06 SHERIDAN COURT 66	FLAT 106	100032093753	FLAT 106	200001408397	1 6

201	FLAT	2	-01 SHERIDAN COURT 66	FLAT 201	100032093764	FLAT 201	200001408398	2 1
202	FLAT	2	-02 SHERIDAN COURT 66	FLAT 202	100032093765	FLAT 202	200001408399	2 2
203	FLAT	2	-03 SHERIDAN COURT 66	FLAT 203	100032093766	FLAT 203	200001408400	2 3
204	FLAT	2	-04 SHERIDAN COURT 66	FLAT 204	100032093767	FLAT 204	200001408401	2 4
205	FLAT	2	-05 SHERIDAN COURT 66	FLAT 205	100032093768	FLAT 205	200001408402	2 5
206	FLAT	2	-06 SHERIDAN COURT 66	FLAT 206	100032093769	FLAT 206	200001408403	2 6
207	FLAT	2	-07 SHERIDAN COURT 66	FLAT 207	100032093770	FLAT 207	200001408404	2 7
208	FLAT	2	-08 SHERIDAN COURT 66	FLAT 208	100032093771	FLAT 208	200001408405	2 8
209	FLAT	2	-09 SHERIDAN COURT 66	FLAT 209	100032093772	FLAT 209	200001408406	2 9
210	FLAT	2	-10 SHERIDAN COURT 66	FLAT 210	100032093773	FLAT 210	200001408407	2 10
211	FLAT	2	-11 SHERIDAN COURT 66	FLAT 211	100032093774	FLAT 211	200001408408	2 11
212	FLAT	2	-12 SHERIDAN COURT 66	FLAT 212	100032093775	FLAT 212	200001408409	2 12
213	FLAT	2	-13 SHERIDAN COURT 66	FLAT 213	100032094008	FLAT 213	200001408410	2 13
214	FLAT	2	-14 SHERIDAN COURT 66	FLAT 214	100032094009	FLAT 214	200001408411	2 14
301	FLAT	3	-01 SHERIDAN COURT 66	FLAT 301	100032094010	FLAT 301	200001408412	3 1
302	FLAT	3	-02 SHERIDAN COURT 66	FLAT 302	100032094011	FLAT 302	200001408413	3 2
303	FLAT	3	-03 SHERIDAN COURT 66	FLAT 303	100032094012	FLAT 303	200001408414	3 3
304	FLAT	3	-04 SHERIDAN COURT 66	FLAT 304	100032094013	FLAT 304	200001408415	3 4
305	FLAT	3	-05 SHERIDAN COURT 66	FLAT 305	100032094014	FLAT 305	200001408416	3 5
306	FLAT	3	-06 SHERIDAN COURT 66	FLAT 306	100032094015	FLAT 306	200001408417	3 6
307	FLAT	3	-07 SHERIDAN COURT 66	FLAT 307	100032094016	FLAT 307	200001408418	3 7
308	FLAT	3	-08 SHERIDAN COURT 66	FLAT 308	100032094017	FLAT 308	200001408419	3 8
309	FLAT	0	FLAT 3-09 SHERIDAN COURT 66	FLAT 309	100032094018	FLAT 309	200001408420	3 9
310	FLAT	3	-10 SHERIDAN COURT 66	FLAT 310	100032094019	FLAT 310	200001408421	3 10
311	FLAT	3	-11 SHERIDAN COURT 66	FLAT 311	100032094020	FLAT 311	200001408422	3 11
312	FLAT	3	-12 SHERIDAN COURT 66	FLAT 312	100032094021	FLAT 312	200001408423	3 12

E.3 Sheridan Court

Sheridan Court is an example of a “tenement” numbered block of flats, that is where the floor and flat number are part of the flat address. Table E.2 shows a selection of the addresses showing the different entries in each file.

In all cases, the floor number is given first and is followed by the flat number. In the case of the NLPG, every property is shown twice with a different UPRN. These are shown in the table as group1 and group 2. Group 1 is largely identified in the data as having a candidate status while group 2 is shown as approved preferred addresses.

There are four different representations of these addresses across the files. Taking the example of flat 5 on floor 3, this will be represented as shown in Table E.4.

Table E.4 Different representations of Flat 305.

Presentation	Files	Comment
FLAT 305	OS-MM NLPG group1	
305 FLAT	Land Registry	This is a minor issue that could be corrected on output
3-05	Council Tax	The floor appears in the number field and the flat in the name field
3-5	NLPG	This implies a number range in this presentation

These different presentations would make interoperability very difficult and are the root cause of two different presentations of the same data occurring in the NLPG.

E.4 Malvern Court

In Malvern court we have an unusual but by no means unique situation where the properties have a valid number with in Malvern Court and an additional valid number on Mapperley Road, the road on which Malvern Court is situated.

Land Registry and OS MasterMap essentially record the properties in the same way. When the different data structures are taken into account this is:

3 Malvern Court, 33 Mapperley Road

The data is shown in Tables E.5 and E.6.

Table E.5 Land Registry records for Malvern Court

REF	SUBBN	BNUM	BNAM	STREET
501567796	1	29	MALVERN COURT	MAPPERLEY ROAD
1002886220	2	31	MALVERN COURT	MAPPERLEY ROAD
501015589	3	33	MALVERN COURT	MAPPERLEY ROAD
501567797	4	35	MALVERN COURT	MAPPERLEY ROAD
501567795	5	37	MALVERN COURT	MAPPERLEY ROAD

Table E.6 OS MasterMap records for Malvern Court

FID	Building Name	Number	STREET
1000002308433704	1 MALVERN COURT	29	MAPPERLEY ROAD
1000002308433706	2 MALVERN COURT	31	MAPPERLEY ROAD
1000002308433708	3 MALVERN COURT	33	MAPPERLEY ROAD
1000002308433710	4 MALVERN COURT	35	MAPPERLEY ROAD
1000002308433712	5 MALVERN COURT	37	MAPPERLEY ROAD

VOA Council Tax ignore ‘Malvern Court’ altogether and relate all addresses directly to Mapperley Road as shown in Table E.7.

Table E.7 VOA Council Tax records for Malvern Court

BAREFNO	Number	STREET
0001520303190B	29	MAPPERLEY ROAD
0001520303192A	31	MAPPERLEY ROAD
00015203033003	33	MAPPERLEY ROAD
00015203035009	35	MAPPERLEY ROAD
00015203037004	37	MAPPERLEY ROAD

The NLPG treats the whole block of Malvern Court (which may be a terrace or mews type of entity) as being 29-37, giving addresses of the form “3 Malvern Court, 29-37 Mapperley Road” as shown in Table E.8.

Table E.8 NLPG records for Malvern Court – Dual Numbered Properties

UPRN	SAON	PAON	STREET	BLPU STAT	LPI STAT
100031570662		33	MAPPERLEY ROAD	1	3
100031570662		29 37 MALVERN COURT	MAPPERLEY ROAD	1	1
100032117363	3	29 37 MALVERN COURT	MAPPERLEY ROAD	1	1
100032117363		33	MAPPERLEY ROAD	1	3
100032117420	1	MALVERN COURT	MAPPERLEY ROAD	1	3
100032117420	1	29 37 MALVERN COURT	MAPPERLEY ROAD	1	1
100032117420		29	MAPPERLEY ROAD	1	3
100032117420		29 MALVERN COURT	MAPPERLEY ROAD	1	3
100032117421	2	31 MALVERN COURT	MAPPERLEY ROAD	1	3
100032117421	2	29 37 MALVERN COURT	MAPPERLEY ROAD	1	1
100032117421		31	MAPPERLEY ROAD	1	3
100032117421		31 MALVERN COURT	MAPPERLEY ROAD	1	3
100032117422	4	29 37 MALVERN COURT	MAPPERLEY ROAD	1	1
100032117422		35	MAPPERLEY ROAD	1	3
100032117423	5	29 37 MALVERN COURT	MAPPERLEY ROAD	1	1
100032117423		37	MAPPERLEY ROAD	1	3

Note. Entries with LPI status '1' are preferred BLPU/LPI combinations

The Council Tax style of presentation is included in NLPG as an alternative address.

A visit was made to Malvern Court and it is in fact a block of five 1920s townhouses with shared drives. In all the files except the council tax file (and these same records carried into the NLPG), Malvern Court is erroneously shown as being a building, which implies some form of multiple occupancy, which is not the case.

There is however one completely incorrect address in the NLPG data that for the UPRN 100031570662 where a single entry is given covering the block of 5 completely separate properties. This implies strongly that Malvern Court is a Block and could cause problems for users.

E.5 Magdala Road

Magdala Road was chosen for study since it has the largest number of flats of any road in Nottingham, with the exception of one with just two tower blocks. The initial desk based analysis showed a significant number of anomalies and therefore a visit was made.

Magdala Road runs off the main road into central Nottingham from the north, the A60. Originally the road would have consisted solely of large late-Victorian detached houses. The style is similar to those areas in midland and northern industrial towns where wealthy business men would have resided.

However over time there has been significant conversion to student-style flats, many cases of backfill, conversion of coach houses into separate dwellings etc. Currently the

area seems to be moving upmarket with many of the properties having been or in the process of being converted into upmarket flats with controlled entry systems on garden gates and post boxes set into outside walls.

It is remarkably difficult to validate the number of flats in some of the buildings since the older conversions are of a style where existing entrances were used in sub-dividing the premise and many of these houses have four or more ground level access doors. In many cases outer gates are locked and access to properties is difficult.

Originally most of these properties were named. Evidence of this exists on the sandstone gateways. However over time, a lot of these names have fallen into disuse.

Issues with Data in the Files

The properties identified in visiting the road and in the data files have been extracted into Excel spreadsheets. The issues are presented here in summary.

Care Homes

Number 15, ASHWELTON HOUSE, is a private residential care home for the elderly, although not identified as such in any of the files. In the Council Tax and NLPG files it is identified wrongly as Welton House. NLPG does also have Ashwelton House but only as a candidate, Welton being the preferred case.

Number 28 is an NCH children's home with 28a an extension to the front of the property being the NCH east midlands regional HQ. Only the NDR and NLPG files identify 28a. No files identify 28 as a Children's home.

Named Properties

All properties are numbered and therefore the use of names is an extra piece of information. However where names are currently used there are three main reasons for wanting them to be in place in a data file:

- Aids interoperability and matching data sets with missing numbers;
- Data redundancy can act as an accuracy check;
- To not use the name can be seen as insulting by the occupant.

However some of the names in the datasets are no longer in use on properties, and some properties now use names when there are none in the data files.

Table E.9 shows the occurrences of named properties in the various source datasets. "✓" indicates use of the name and "x" indicates the use of a wrong name, suggesting that the use of names in the file in no way matches what can be observed on the ground.

Table E.9. Occurrences of property names

Name	Number	Survey	MM	CT	LR	NLPG
HEATHERLANDS	1A			✓		✓
ASHLING HOUSE	3			✓		✓
DRIFT HOUSE	9					✓
WELTON HOUSE	9			x		x
ASHWELTON HOUSE	15	✓				✓
MEADFOOT HOUSE	16	✓				
ABBEYFIELD HOUSE	17					✓
ASHDALE HOUSE	18	✓				
ATHOLL HOUSE	20	✓				
RED LODGE				✓		✓
THE CHESTNUTS	25	✓				
MAGNOLIA HOUSE	25A	✓				
THANASSIS	27	✓				
MAGDALA MOUNT	29	✓		✓		✓
BEULAH HOUSE	35	✓				
WESTFIELD HOUSE	42	✓				
THE LIMES	44	✓			✓	

Divergence between files

One of the main reasons for visiting Magdala Road was to identify why there was divergence between the files in the recording of internal breakdowns of properties. Actually undertaking a survey illustrated the difficulties in this.

Number 38 for example has been split into 10 flats numbered 1-2 and A-H, there are a series of mail boxes set into the outside wall and these have been confusingly numbered so it appears at first sight as if numbers 38, 38A and 38B are separate properties the reality is different.

Table E.10 shows what was found on the ground and what was in the files. The serious issue here seems to be the creation of entities in all of the files that don't exist on the ground it is the case that this particular property is very confusing.

Table E.10. Occurrences of flats

		Survey	MM	CT	LR	NLPG
	38		✓		✓	✓
	38A		✓		✓	✓
FLAT 1	38	✓				
FLAT 1	38A					✓
FLAT 2	38A			✓		✓
FLAT 2	38	✓				
FLAT 3	38A	✓		✓		✓
FLAT A	38	✓		✓		✓
FLAT B	38	✓		✓		✓
FLAT C	38	✓		✓		✓
FLAT D	38	✓		✓		✓
FLAT E	38	✓		✓		✓
FLAT F	38	✓		✓		✓
FLAT G	38	✓		✓		✓
FLAT H	38	✓		✓		✓
FLAT I	38					✓
FLAT J	38					✓

Other Issues

Missing Properties

All the files missed some properties, in some cases files identified many more flats than were evident on the ground, but these may be only identifiable internally.

Not keeping up with Changes

This is a fluid area. Some property has been demolished and replaced. In the case of number 22, all files identify it as a single property with flat(s) it is in fact now a pair of fairly new semis numbered 22A and 22B

Duplicated properties

In this case the Land Registry data had a significant number of duplicates with different unique keys.

Errors retained in files

There are a number of occasions where the NLPG retains errors by using a BLPU status of 8 instead of 9. Examples are 31A where accidentally created duplicates are flagged as historic rather than errors (status 9). This also happens with Park Mews where this been reclassified as a road.

E.6 Newcastle Chambers

This example was chosen to illustrate the difficulties in comparing business addresses between the files. The entries on NLPG, Land Registry, OS MasterMap and VOA Non-domestic rates are shown in Tables E.11 – E.14.

Table E.11 NLPG records

UPRN	SAON	PAON	STREET	BLPU Status	LPI Status
100032095094	1	NEWCASTLE CHAMBERS	ANGEL ROW	5	5
100032095094		1	NEWCASTLE CHAMBERS	5	1
100032095095		2	NEWCASTLE CHAMBERS	1	1
100032095096		3	NEWCASTLE CHAMBERS	1	1
200001401790	1 UNIT A	NEWCASTLE CHAMBERS	ANGEL ROW	1	1
200001401791	UNIT B1	NEWCASTLE CHAMBERS	ANGEL ROW	1	1
200001401792	1 UNIT C	NEWCASTLE CHAMBERS	ANGEL ROW	1	1
200001401793	UNIT E1	NEWCASTLE CHAMBERS	ANGEL ROW	1	1
200001401794	BASEMENT ROOM 58	NEWCASTLE CHAMBERS	ANGEL ROW	1	1
200001401795	BASEMENT ROOM 59	NEWCASTLE CHAMBERS	ANGEL ROW	1	1
200001401796	BASEMENT ROOM 60	NEWCASTLE CHAMBERS	ANGEL ROW	1	1
200001401797	BASEMENT ROOM 61	NEWCASTLE CHAMBERS	ANGEL ROW	1	1
200001401798	BASEMENT ROOM 62	NEWCASTLE CHAMBERS	ANGEL ROW	1	1

Table E.12 Land Registry records

REF	SUB NO	SUB NAME	BUILDING NAME	STREET
501011767	1		NEWCASTLE CHAMBERS	ANGEL ROW
501011768			NEWCASTLE CHAMBERS	ANGEL ROW
501011769	3		NEWCASTLE CHAMBERS	ANGEL ROW
1001712355	2	SOUTH WING	NEWCASTLE CHAMBERS	ANGEL ROW

Table E.13 OS MasterMap records

FID	ORGANISATION	BUILDING NAME	STREET
1000002308426518	AUDIO VISUAL ARTS LTD	1 NEWCASTLE CHAMBERS	ANGEL ROW
1000002308426519	DICKINSONS PROPERTY CONSULTANTS	NEWCASTLE CHAMBERS	ANGEL ROW
1000002308426520	OPEN DOOR	3 NEWCASTLE CHAMBERS	ANGEL ROW
1000002308919187		3 NEWCASTLE CHAMBERS	ANGEL ROW

Table E.14 VOA NDR records

BAREFNO	NONSDISC	No	NAME	STREET
N00013227000025	OFFICE AND PREMISES	0	UNIT A 1 NEWCASTLE CHAMBERS	ANGEL ROW
N00013227000237	OFFICE AND PREMISES	0	UNIT C 1 NEWCASTLE CHAMBERS	ANGEL ROW
N00013227000044		0	UNIT B 1 NEWCASTLE CHAMBERS	ANGEL ROW
N00013227000256	OFFICE AND PREMISES	0	UNIT E 1 NEWCASTLE CHAMBERS	ANGEL ROW
N00013227000082		3	NEWCASTLE CHAMBERS	ANGEL ROW

As in the other examples, the NLPG has more records than any of the other files. However this is a multiple occupancy office/retail building and a search on Google shows a number 25 and a number of trading name entities that are not in any of the files. This suggests that there are more occupiers than shown here.

Whilst most of the records can be matched between some of the files, only number 3 matches in all the files and even this case is problematic because of the Open Door, 3 Newcastle Chambers entry in the OS MasterMap data that suggests that 3 may be a divided property.

Other issues in the data are the movement of Newcastle Chambers into the street field on three of the NLPG records, the question of whether UNIT B is a sub-division of number 1 as shown in the Council Tax file or B1 as shown in the NLPG or possibly neither of these.

A survey identified that Newcastle Chambers is an alleyway between two retail premises on Angel Row. Existing occupants are

Doors to East side of alley:

Dickinson's Property Consultants UNIT C
 Work tribe UNIT E
 Martin Nouch (chartered surveyor)

Doors to West side of Alley:

W8 (recording studio)
 Bush Doctor (West Indian Café)
 Vani (?)

In addition, there were two other doors with bells, with no indication of occupant, that were not rear accesses to the retail premises.

There is no evidence to suggest that any of the files are wrong in what they have represented, although it is clear that all of them have only part of the story.

E.7 OS MasterMap commercial multiple occupancy

Table E.15 shows the OS MasterMap multiple-occupancy business addresses for Nottingham. For each entry, the FeatureID, name of the business, postcode and number of units are given. The main point of interest is how few multiple occupations are identified in relation to the number that should be identified. Some 35 entries have multiple occupancy counts and are also identified by entries in the organisational name field as organisations.

Table E.15 OS MasterMap Multiple Occupancy Business Addresses

FID	NAME	Postcode	No	comment
1000002308426225	PREMIER HOUSE	NG1 2NA	40	Serviced Offices
1000002307023212	SHOWMANS OVERWINTERING GROUND	NG8 5FH	20	
1000002308426934	DAVID SAIL ,EAST MIDLANDS REGIONAL OFFICE	NG1 6BJ	11	Estate agent
1000002308316000	SPREAD EAGLE	NG7 3PE	9	Pub
1000002308474214	POINTBIO PLC	NG2 4DS	7	
1000002308366777	1 STOP PROMOTIONS	NG1 5BB	6	
1000002308535530	FOSTERS MOBILE WORKSHOP	NG2 4BG	6	
1000002308429913	NOTTINGHAM CITY COUNCIL	NG1 4BT	6	Government
1000002308365959	MAIL BOXES ETC	NG1 5FQ	5	
1000002308919168	THE BUSINESS CENTRE	NG1 1LH	4	College
1000002308534433	PARK YACHT INN	NG2 4DS	4	PUB
1000002308433832	DENTAL SURGERY	NG3 5AA	4	Dentist
1000002308433862	HIMALAYA TANDOORI RESTAURANT	NG1 3FT	4	Restaurant
1000002308362493	DENTAL SURGERY	NG1 5DN	3	Dentist
1000002307023361	BEE ENTERPRISES WORKSHOP	NG8 6GD	3	
1000002308427084	BENEFITS AGENCY	NG2 3BZ	3	Government
1000002308476115	TERRYS BUTCHERS	NG2 4LW	3	Shop
1000002308430377	BENEFITS AGENCY	NG1 6AX	3	Government
1000002308316310	IKANO FINANCIAL SERVICES LTD	NG7 5LN	2	
1000002308312143	DENTAL SURGERY	NG7 1PZ	2	Dentist
1000002308162070	DIRECT VALUATIONS	NG9 2NX	2	
1000002308362936	EUROPADISC	NG1 6JE	2	
1000002308323648	VISTA SIGNS	NG7 7DA	2	
1000002308324009	MARK ONEDENTAL SURGERY	NG7 7DA	2	Dentist
1000002308935929	HORSE AND GROOM	NG7 7EA	2	PUB
1000002308480928	M M CLOTHING	NG3 1JE	2	
1000002308476111	CORAL	NG2 4LW	2	Shop
1000002308476112	NAIKS NEWSAGENT	NG2 4LW	2	Shop
1000002308536222	TIGER EUROPEAN	NG2 4AP	2	
1000002308536295	JOHNSONS BAKERS	NG2 4DE	2	Shop
1000002308536236	MAID MARIAN STORES	NG2 4AX	2	Shop
1000002308939354	LIGHT YEARS LTD	NG2 4AX	2	
1000002308423146	P T STORES	NG2 2LH	2	Shop
1000002308430375	INDEPENDENT TRIBUNAL SERVICE	NG1 6AZ	2	
1000002308438396	EL SOMBRERO	NG5 2FS	2	Restaurant

The comment field includes a description of the property that has been obtained either from the address or by research what is surprising is the proportion of dentists identified as being in multiple occupation, however it is not clear whether this is multiple dentist practises sharing the same premise or simply a surgery with a flat above.

E.8 Banks

It became clear when examining the data that there were differences in the manner that each of the Acacia partners handled business names. It is difficult in general to determine what the correct name for a business is without asking. However banks are a group of organisations who clearly identify both their trading names and their legal identities on their Web sites. For this reason it was decided to use banks of a test of the correct recording of business names and the comparability between files. To facilitate this, the trading and legal names of all of the banks trading in this area were obtained.

To allow the consideration of similar areas firstly a set of banks were extracted from the NDR file then matching banks were extracted from the NLPG and OS MasterMap data based on postcodes, no matching named banks were found in these areas in the Land Registry data.

The most surprising result of this analysis was that a very large proportion of the names were wrong, referring to a previous name of the bank, for example “Midland”, “Trustees Savings Bank” and even “The Westminster Bank”.

In general there was little to choose in terms of correctness and accuracy between the NLPG and NDR files with both predominantly using names that were neither the current trading nor the current legal names of the banks concerned. The OS MasterMap data was generally better showing sign of some form of standardisation process. However, even here there was a lack of consistency, with for some banks the legal name being used and some the trading name, though for each bank a choice of one or other seems to have been made.

In general, if a name is required either for definition or business purposes it would generally be better to use the trading name since that is what will be displayed over the premise and it is not always clear from a legal name what the trading entity might be, e.g. “HBOS PLC” is the legal name of both “The Halifax” and “Bank of Scotland” bank branches.

Abbreviations

BLPU	basic land and property unit
CTAX	Council Tax
ETL	extract, transform, load
LLPG	local land and property gazetteer
LPI	land and property indicator
NDR	non-domestic rates
NLPG	National Land and Property Gazetteer
PAF	Postcode Address File
PAO	primary addressable object
PO	Post Office
SAO	secondary addressable object
TOID	topographic identifier
UPRN	unique property reference number
VOA	Valuation Office Agency