UCL DEPARTMENT OF CIVIL,
ENVIRONMENTAL & GEOMATIC ENGINEERING

"DESIGNING AN APPROACH FOR AUTOMATIC DATA QUALITY CHECKING AND REPORTING BASED ON METADATA"

By Sofia Covarrubias Zabala

Supervisor: Dr. Claire Ellul

MSc GIS

Department of Civil, Environmental and Geomatic Engineering

University College London

London, September 12th 2018

Abstract:

Logical consistency of spatial features is one of the data quality challenges and the traditional way to handle it is through the definition of spatial integrity constraints to be stored as part of the product specifications and implemented in the database to prevent erroneous data entry. However, this approach is not applicable to certain organizations, because it requires technical knowledge which is not always available at the required scale.

Based on the case of the London Borough of Hackney, this research explores the use of spatial constraints stored in a metadata repository, to create potentially automatable quality check routines and reports, using FME. Specifically, the research tries to answer if a process can be defined to facilitate the automation of spatial data quality rules validation based on metadata about spatial constraints.

A qualitative approach was used to help pilot data owners express their constraints in natural language which were then translated into a formalized structure, tested in FME, and reported to the data owners to receive their feedback.

The results proved the formalized structure was suitable for basic automatic checks but not enough to preserve the communicational value of the natural language, thus a combined approach was proposed. However, this approach is limited by the fact that, while it does not require scripting language, it needs certain knowledge about standardized terminology to refer to spatial relations in a machine-readable format, which cannot be assumed as universally understood.

Furthermore, the results showed that these basic reports can be improved through the distinction of types of issues and it was possible to identify a suitable approach to handle some specific kind of issues (coordinate precision problems and lack of snapping tools to define features). However, the cases analysed in detail were not enough to determine a general model for these detailed checks beyond those two particular cases, and suggest, on the contrary, that the specific workflow for detailed analyses is strongly dependant on each case.

Acknowledgements:

This work would not had been possible without the support of the London Borough of Hackney. I would like to thank specially to the GIS team, Sandrine Balley, Marta Villalobos and Tapan Perkins, for their patient guidance, generous sharing of invaluable knowledge, experience and time, and for the warmth with which they received me in their team. Furthermore, I am also grateful to each of the data owners who kindly accepted to anonymously participate in this process. I hope this research will be a small contribution to the great work you do each day.

I would like as well to extend my gratitude to my supervisor Claire Ellul and Nikolaos Papapesios for their invaluable advice, and continued support throughout the project.

Finally, I would like to thank my family, particularly my husband Rodrigo for his unconditional support and my mother Gabriela.

Table of Contents

List of figures

List of tables

Abbreviations:

9IM	Nine Intersection Model
AGI	Association for Geographic Information
BSI	British Standards Institution
DE9IM	Dimensionally extended nine intersection model
DO	Data Owner
GISc	Geographical Information Science
INSPIRE	INfrastructure for SPatial InfoRmation in the European Community
ISO	International Organization for Standardization
OGC	Open Geospatial Consortium
OSMM	Ordnance Survey Master Map
RDBMS	Relational Database Management System
SDI	Spatial Data Infrastructure (see glossary for definition)
SQL	Structured Query Language
UML	Unified Modelling Language

Glossary:

TERM	DEFINITION
Feature	A representation of a real-world object on a map (ESRI, s.f.).
Topology	The relative location of geographic phenomena independent of their exact position (Smith, et al., 2018)
Spatial data	Information about the location and shape of geographic features and the relationships between them, usually stored as coordinates and topology. Any data that can be mapped. (ESRI, s.f.)
Spatial database	A database is a collection of data used to represent information of interest to an information system. A spatial database is the one that includes spatial data (Atzeni, et al., 1999)
Universe of discourse	View of the real or hypothetical world that includes everything of interest (BSI, 2015)
Unified Modelling Language	The Unified Modelling Language (UML) is a general-purpose, developmental, modelling language in the field of software engineering that is intended to provide a standard way to visualize the design of a system. (Addison-Wesley, 2005)
Data warehouse	A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users. (Oracle, s.f.)
Cardinality	The cardinality of a binary topological relation specifies the number of objects (if any) with which an object can have the specified relation (Vallieres, et al., 2006).
Earthlight	London Borough of Hackney's intranet mapping application

Chapter 1: Introduction

The current times had been called the information age and in it, information is the main asset of companies, governments and individuals, as the base of decision making (Castells, 2010). Most of this information has a location as, for how Longley et al. (2015) said, "almost everything that happens, happens somewhere" and the location of things, people or events affects the possible relations between them. To include this location when storing data makes what is known as spatial data.

Geographical information science had allowed the production, storage, share and analysis of these spatial data, and its analysis underlie decision making in diverse matters. In this context, the quality of the spatial data used for decision making should be assured by data managers, but the spatial characteristics of geographic information (form and location) is source of specific kind of quality issues and management challenges that have made it a research field by itself, and a topic of special interest for the Infrastructure for Spatial Information in the European Community (INSPIRE), a legal and technical framework for sharing and reusing spatial data in Europe, to which the UK adhere.

Following ISO 9000, INSPIRE understands quality as the "degree to which a set of inherent characteristics fulfils requirements" (JRC European Commission, 2013). In relational database management system (RDBMS) terminology, these requirements include spatial integrity constraints which rule how the data should be and are defined as the "conditions that must always be valid for the model of interest" (Cockcroft, 1997). The standard way to handle them in RDBMS is as part of the product specifications and database design, storing them using the Unified Modelling Language (UML) and implementing them in the database using a scripting language such as the Structured Query Language (SQL).

But the previous approach seems unrealistic for certain type or organizations such as councils, as the London Borough of Hackney (LBH). On the one hand, council's spatial data are big, complex and dynamic, with a wide range of people with different profiles involved in the database creation and maintenance process, making a formal specification approach very difficult and with high risks of non-adoption. At the same time, LBH's personnel with the required expertise is not enough to take care of this task for all LBH's databases (more than 100), neither to maintain these constraints in case of changes, while from a data governance point of view, is not even recommendable.

Additionally, this way to handle constraints as part of the design of the database has the drawbacks that it makes them hidden in the codes so are difficult to maintain (Van Oosterom, 2006) and doesn't

have a communicational value to make the data users, such as managers and editors, aware of the data formal requirements.

One of the constraints management approaches that has been proposed in literature, suggested by Cockcroft (2004) and Van Oosterom (2006), seems especially useful in this context: storing the constraints in a metadata repository, from where they can be automatically checked or implemented in the subsystems of the infrastructure. As both researchers assert, this approach is more accessible as the constraints are not hidden in the database scripting code, useful for communication purposes, easier to maintain and system-independent so easier in case of system migration.

LBH's GIS team has implemented an extensive spatial data infrastructure and a metadata model including spatial constraints, which is coupled with FME for data management (see *Chapter 2* for further explanation about FME). In this context, the aim of this research was to explore the use of metadata about spatial constraints to create potentially automatable quality check routines and reports, based on the case of the LBH. Specifically, the research tries to answer if a process can be defined to facilitate the automation of spatial data quality (DQ) rules validation based on metadata about spatial constraints, with the following sub-questions: 1) How data owners (DOs) currently express the constraints in the database and what can we learn from this about how to communicate them to other users? 2) Which is the best way to store topological constraints as metadata with the double purpose of being machine readable and communicate to other users how the data is expected to be? 3) Is it possible to define a general model of quality check routines based on topological constraints using FME? and 4) How should the results of a quality check routine be reported to the relevant DOs to be effective for the purpose of improving DQ?

The scope was limited to 2D spatial data in vector format, and the kind of constraints analysed were those related to topological relationships between features, as were selected by the LBH as the focus of this project. The topological consistency of single features is not in the scope of this project.

The rest of this report was divided as follows: the second chapter presents the LBH's spatial data infrastructure. The third chapter reviews the main literature that supports the work. The fourth chapter explains the methodology used. The fifth chapter presents the results and main findings, which are discussed in the sixth chapter. Finally, a conclusion chapter sums up and explains the meaning of the concluding remarks, while presents some wider implications and further work needed.

Chapter 2: LBH's data management system

LBH's Spatial Data Infrastructure (SDI) integrates data from the different local services (such as Education, Planning, Housing and Social Services) and data provided by external entities such as the Ordnance Survey, the Land Registry, the National Health System and Transport for London. All these data are stored in Geostore, LBH's corporate spatial data warehouse, and is populated and visualized by the services using an Intranet mapping application called Earthlight. All the layers published in Earthlight come from live data from Geostore, so when a layer gets updated by a service area (using Earthlight or another GIS software) the changes are instantly reflected in all Earthlight maps containing this layer, facilitating data sharing among the council. In Earthlight, each layer is also associated to a metadata record to facilitate discovery, exploration and exploitation. The metadata model is based on the Gemini2 UK standard, enriched to fit the council's internal requirements.

For data management, the system is coupled with FME (Feature Manipulation Engine, developed by Safe Software), the market-leading spatial data integration platform. FME basically works with an interface to create workflows that go from data reading (extraction from its source) to data writing (loading into the destination) passing through whatever the user wants to do to the data, which can be a vast range of tools including data conversion, data transformation (regarding content or structure) or data validation tools. Furthermore, FME does not require a programming language, has automation capabilities and the stored workflows can be re-run automatically, according to a schedule or in response to a trigger (Safe Software, n.d.), which make it suitable for the design of automatable quality routines based on metadata.

The general design, management and support of the SDI are duties of the corporate GIS team formed by three GIS experts, while the data itself is managed by each service as part of a data governance framework. For this purpose, each service defines a "data owner" (DO) per certain number of datasets (depending on the complexity of them), whose role is to have well-specified datasets, enabling high DQ and publishing his/her datasets in a usable format with relevant metadata. This role is complemented by several "data custodians" in charge of the maintenance of the data (creating, updating or deleting records).

As the corporate GIS lead explains (S Balley 2018, personal communication, 9 August), DOs have various profiles, and owning data is not their main duty. There is also a relatively high turnover. This limits the possibilities of having formal specifications across all the services following the international standards recommendations. However, from a data governance point of view, it is considered healthy to set someone in each service as responsible for the data documentation even if they are not

specialists. It ensures that services areas are aware of basic data management concepts and, with the support of the GIS corporate team, may raise the chances of having high-quality datasets. In this context, brief and user-friendly metadata including spatial constraints is considered the maximum they can ask data owners and the right level of detail and abstraction that the rest of the users in and out Hackney need.

When the DO populates the metadata of a dataset in the system, he/she has to define the integrity constraints of it to ensure DQ, i.e. describe the rules the data is supposed to comply with, regarding attributes and geometry. This information is meant to serve as a guideline to the data custodians to know how the data is expected to be. These constraints can be either internal to the dataset or related to other dataset and currently should be filled directly as text without any kind of predefined format.

Chapter 3: Literature review

In order to provide a theoretical underpinning for the work, this chapter presents a review of the theory behind the concept of DQ in GISc, with particular focus on producer and user perspectives (as the LBH teams consists of both). Constraints –the rules used to ensure that a dataset is captured to a specific level– are then described as a tool to ensure DQ, with specific focus on those relating to topological relationships between features. The specific scope of the study –topological relationships– are described then, and previous attempts at the formal encoding of topological rules are reviewed to determine if a suitable methodology to carry out this task can be developed from existing work. Metadata, which is used to document the quality of a dataset in terms of the conformity to predefined rules, and (theoretically) allow users such as the LBH's services to assess its fitness for their task, is then described, in particular within the context of the INSPIRE project which is mandated for many datasets at LBH.

3.1. Data quality

Since geospatial data is a representation of reality –a model which is never perfect–, the concepts of *error* and *quality* are a constituent part of it. As Chrisman (2006) explains, in its origins spatial DQ was focused firstly on its positional accuracy (i.e. the difference between the location of a feature in the map and in reality) and secondly in the topological consistency of the features (i.e. that a polygon is properly closed so, for example, its colour does not spread all over the map[1]). But the concept evolved and nowadays the standard definitions are much wider, based on the concept of *fitness for purpose* where the producer informs the results of quality tests (via metadata) and it is the user who judges, based on that information, if the quality is enough for his/her purposes (Chrisman, 2006). This is the case of the definition used by INSPIRE for whom quality corresponds to the "degree to which a set of inherent characteristics fulfils requirements" (JRC European Commission, 2013) and the reason why metadata is so relevant for INSPIRE.

3.1.1. User and Producer Perspectives on data quality

The current ISO standard on geographic information DQ, ISO 19157 distinguishes DQ from the perspective of the producer and from the perspective of an external user (*Figure 1*). As data producers and data users may have different requirements, the quality of a dataset depends on who is evaluating

[1] For clarification, even though the term –topological– is the same, here it refers to the geometry of single features, while this project is focused on topological relations between two features, namely, *binary topological relations*.

it. For the data producer, DQ corresponds to the difference between a dataset and its *universe of discourse* (i.e. the perfect dataset that corresponds to the data product specification) (BSI, 2018). The user, instead, defines his/her universe of discourse according to his/her requirements for whether to use a dataset or not (BSI, 2018). Related to these, the concepts of internal quality (focussing on the level of error in the data in its attempt to represent the real world) and external quality (defined in terms of user needs) are also frequently used (Devillers & Jeansoulin, 2006).

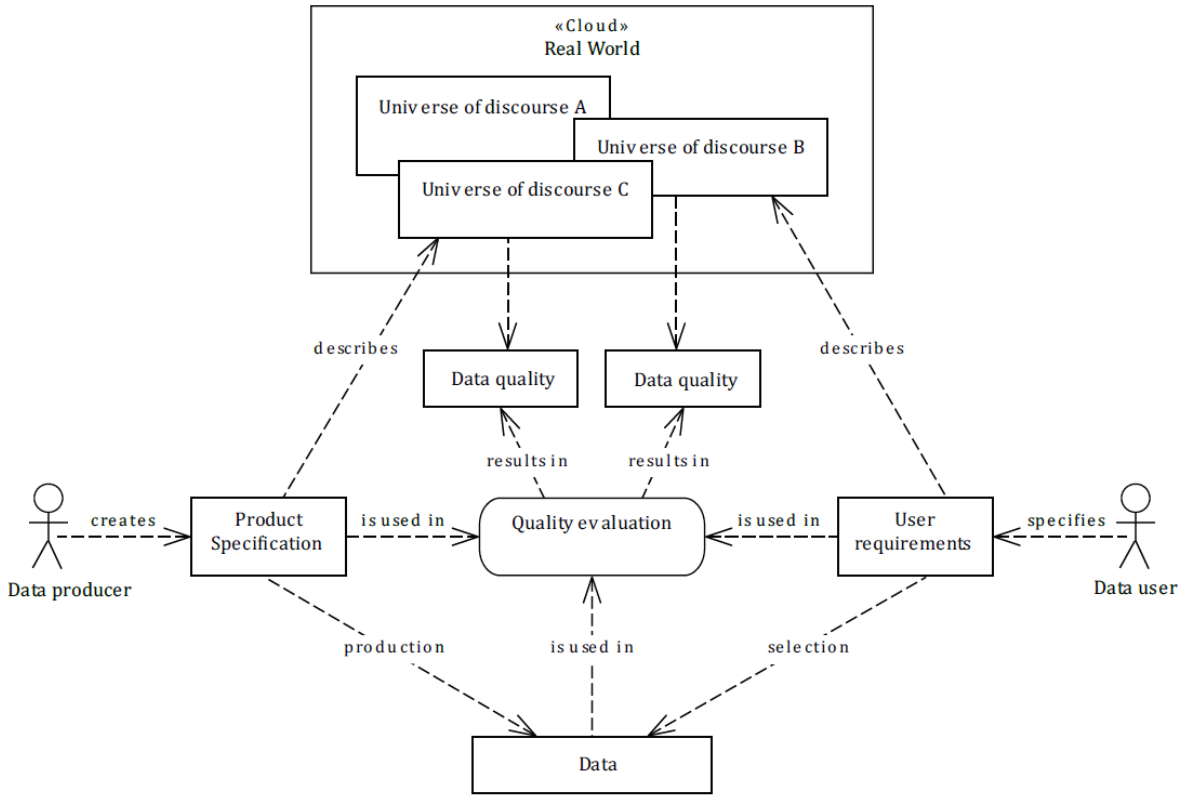


Figure 1: Framework of DQ concepts, defined in ISO 19157 (BSI, 2018, p. 29) *and used by INSPIRE. In it, DQ is not an objective threshold but a subjective evaluation according to who is evaluating.*

3.1.2. Data Quality Elements

According to this ISO standard and to INSPIRE, DQ shall be assessed using established DQ elements that allow evaluating the difference between the dataset and the universe of discourse of a) the data product specifications or b) the user requirements (BSI, 2006; JRC European Commission, 2013). *Table 1* provides the definition of all the quality elements and sub-element considered by ISO 19157 to assess DQ.

Table 1: DQ elements and sub-elements on ISO 19157 and INSIRE

DQ element and sub-elements	Description
1. Completeness	Presence and absence of features, their attributes and relationships.
• Commission	Excess data present in a dataset.
• Omission	Data absent from a dataset.
2. Logical consistency	Degree of adherence to logical rules of data structure, attribution and relationships.
• Conceptual consistency	Adherence to rules of the conceptual schema;
• Domain consistency	Adherence of values to the value domains.
• Format consistency	Degree to which data are stored in accordance with the physical structure of the dataset.
• Topological consistency*	Correctness of the explicitly encoded topological characteristics of a dataset.
3. Positional accuracy	Accuracy of the position of features within a spatial reference system.
• Absolute or external accuracy	Closeness of reported coordinate values to values accepted as or being true.
• Relative or internal accuracy	Closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true.
• Gridded data positional accuracy	Closeness of gridded data spatial position values to values accepted as or being true.
4. Thematic accuracy	Accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classification of features and their relationships.
• Classification correctness	Comparison of the classes assigned to features or their attributes to a universe of discourse
• Non-quantitative attribute correctness	Measure of whether a non-quantitative attribute is correct or incorrect.
• Quantitative attribute accuracy	Closeness of the value of a quantitative attribute to a value accepted as or known to be true.
5. Temporal quality	Quality of the temporal attributes and temporal relationships of features.
• Accuracy of a time measurement	Closeness of reported time measurements to values accepted as or known to be true.
• Temporal consistency	Correctness of the order of events.
• Temporal validity	Validity of data with respect to time.
6. Usability element	Usability evaluation may be based on specific user requirements that cannot be described using the quality elements described above. In this case, the usability element shall be used to describe specific quality information about a dataset's suitability for a particular application or conformance to a set of requirements.

Source: ISO 19157 – Geographic Information – Data Quality (BSI, 2018)
* This sub-element refers to the geometry of single features (e.g. that a polygon should be closed) while this project is focused on topological relations between two features (binary topological relations) which should be defined in the conceptual schema of a RDMS and, therefore, relate to the quality sub-element "conceptual consistency".

3.2. Improving Data Quality – Constraints

In database management systems, integrity constraints –or rules– are "conditions that must always be valid for the model of interest" (Cockcroft, 1997), defining the rules limiting the values that can be input into a dataset and check that values between related datasets are consistent. Statements such

as “the gender should be expressed as *F* or *M*” or “a date should be stored in the format *DD/MMM/YYYY*” are examples of integrity constraints where compliance ensures that a high-quality dataset is created. In the case of spatial databases, spatial constraints deal with the consistency of the geographic properties of the data.

Based on the classification made by Cockcroft (1997) and refined by Van Oosterom (2006), spatial integrity constraints relate to the spatial properties of the object (its size or shape) or are derived from the spatial relations between objects (*Figure 2*). The latter can be subdivided in constraints related to the topology of the relation, its direction, distance or a mixture of them, with topological constraints being the focus of this research.

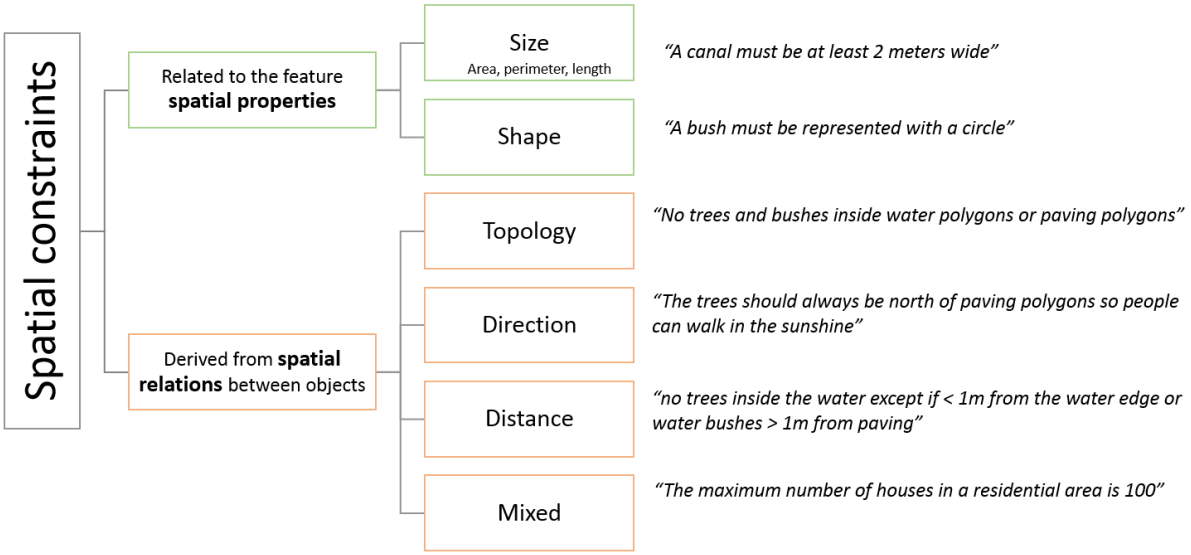


Figure 2: Classification of spatial constraints. Based on the work of Van Oosterom (2006)

In particular, topological constraints are rules related to topological properties and spatial relations (see *Section 3.3*), and they should be constructed using a neighbourhood and containment framework (Van Oosterom, 2006). Smith et al. (2018) state that topology refers to “the relative location of geographic phenomena independent of their exact position” and a property is topological “if it survives to stretching and distorting of space” (Smith, et al., 2018). From this definition, topology thus refers to the relationships between things. Examples of topological properties are the adjacency of two land parcels or the containment of a point within an area and topological constraints in a 2D context are, for example: “a road cannot cross waterbodies” or “a street lighting cannot be within a building”.

However, the data acquisition, data entry or update, or the integration of different data sources can produce errors that may violate these rules, generating differences between how the data should be according to the producer and how the data actually is. Therefore, the validation of these spatial

constraints had been proven as a tool to improve DQ which allows to identify inconsistencies or to prevent them (Borges, et al., 2002; Vallieres, et al., 2006; Van Oosterom, 2006).

3.3. Defining topological constraints

The process to define spatial constraints and, in particular, topological constraints, is more difficult than other constraints due to special characteristics of geographic data (Borges, et al., 2002). This process can be divided into 2 steps (Van Oosterom, 2006): first, their expression in natural language (i.e. how the user would talk about them in daily life –see *Section 5.1* for related experiments–) and second, their translation using a mathematical formalization of spatial relations and a machine-readable structure that ensures the complete definition of the topological rule. Additionally, it may be required to express them in a language readable by the specific system in which it is wanted to be implemented, but, for this research, the objective is to explore the use of constraints expressed in a general system-free form.

Sub-section 3.3.1 presents the accepted framework for the mathematical formalization of spatial relationships, along with the linguistic predicates that had been defined to refer to the fundamental set of spatial relations and its limitations. *Sub-section 3.3.2* presents how a machine-readable structure can organise all the elements required to properly define a topological constraint (including but not limited to the spatial relation predicate(s)).

3.3.1. Mathematical formalization of topological constraints and its linguistic predicates

The accepted framework for the formalization of spatial relationships, according to ISO 19125 (BSI, 2006), is the one developed by Egenhofer and Herring (1994) and extended by Clementini and Di Felice (1996). Egenhofer and Herring proposed the so-called 9-Intersection Model (9IM) to describe the relationship between two geometric objects in a two-dimensional space, based on the intersection between their boundaries, interiors and exteriors (See *Figure 3* for their geometric definition). This relation is described using a 9-intersection matrix which specifies whether each intersection is empty or non-empty. Clementini and Di Felice's extension specifies the dimension of a non-empty intersection. This extension was called the Dimensionally Extended 9 Intersection Model (DE9IM). In the matrix (*Figure 4*), each cell should be filled with the values {T, F, *, 0, 1 ,2}, where "T" means that there is an intersection with any dimension, "F" means that the intersection is empty, "*" that it can be empty or non-empty, "0" that it corresponds to a point, "1" that it corresponds to a line, and "2" that it corresponds to a polygon (Vallieres, et al., 2006).

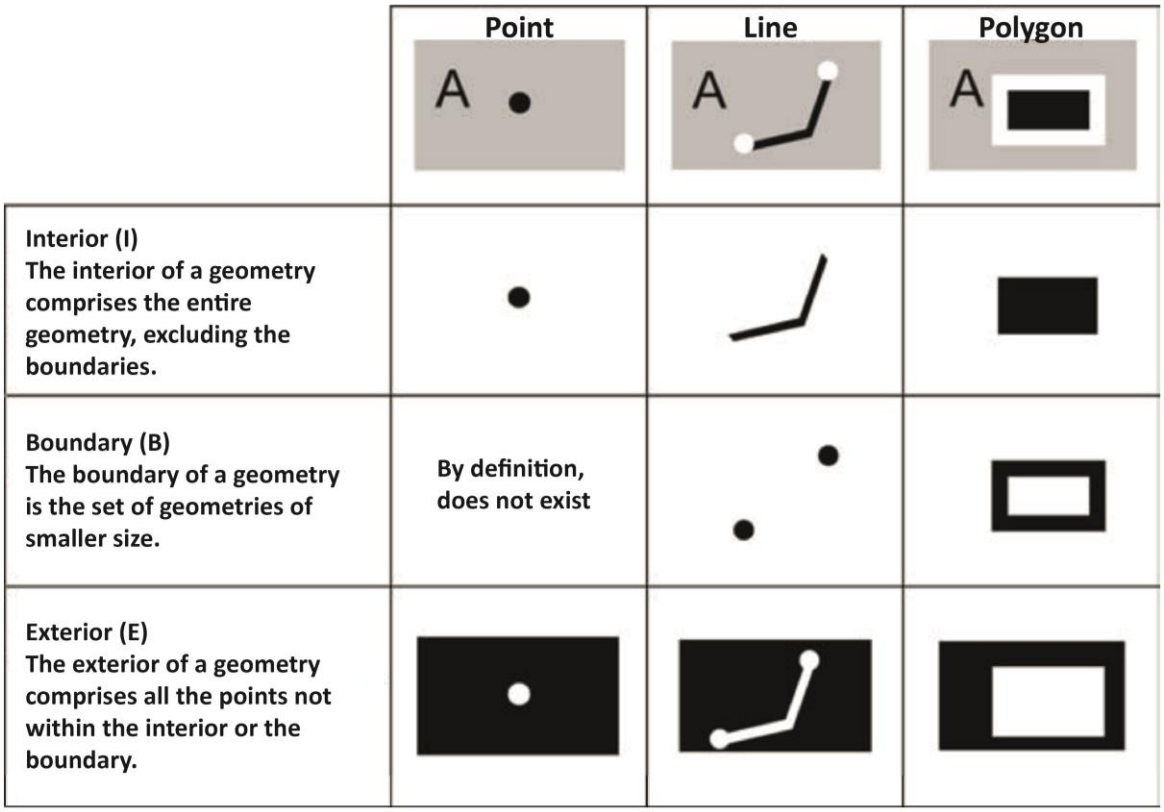


Figure 3: Concepts of Interior, boundary and exterior for each geometric primitive (Vallieres, et al., 2006)

	Interior	Boundary	Exterior
Interior	$dim(I(a) \cap I(b))$	$dim(I(a) \cap B(b))$	$dim(I(a) \cap E(b))$
Boundary	$dim(B(a) \cap I(b))$	$dim(B(a) \cap B(b))$	$dim(B(a) \cap E(b))$
Exterior	$dim(E(a) \cap I(b))$	$dim(E(a) \cap B(b))$	$dim(E(a) \cap E(b))$

Figure 4: The DE9IM (BSI, 2006). *"Dim" stands for dimension, "I" for interior, "B" for boundary and "E" for exterior. Each cell should be filled with the values {T, F, *, 0, 1, 2}.* (Vallieres, et al., 2006).

The categorization of the 81 possible relations according to the 9IM, and the use of named predicates to refer to subsets of the possible relations has been studied since the 9IM was proposed (Mark & Egenhofer, 1995) as a way to help the definition of spatial relations. While it had been shown that it is possible to categorize the binary relations in a fundamental set of spatial relations using the 9IM (Egenhofer & Herring, 1994), Mark and Hegenhofer states that there is not a single set of spatial linguistic predicates able to describe all the spatial relations in an exhaustive and mutually exclusive way for the human understanding, as it is possible to do mathematically with the 9IM (1994).

As an attempt to define a common terminology to refer to topological relations in GIS, the International Organization for Standardization (ISO) and the Open Geospatial Consortium (OGC) set names for subsets of relations –disjoint, touches, crosses, within, overlaps, contains, intersects and equals– (BSI, 2006; OGC, 2011). *Figure 5* shows the 8 named predicates defined by ISO and OGC and its mathematical definition using the DE9IM.

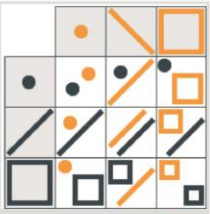
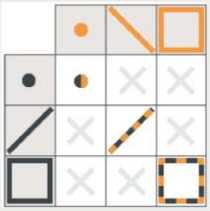
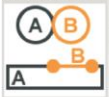
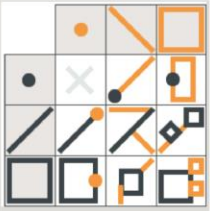
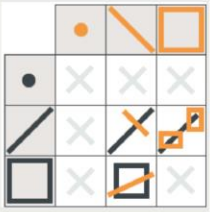
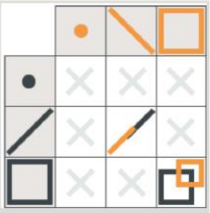
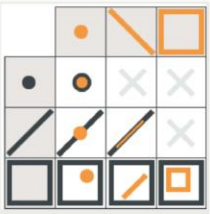
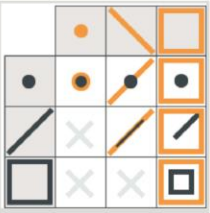
Predicate	Description	Geometry Examples	Pattern Matrix*
Intersects	The two features are not **disjoint**, as defined next.		
Disjoint 	The boundaries and interiors do not intersect.		F F * F F * * * *
Equals 	The features have the same boundary and the same interior.		T * F * * F F F *
Touches 	The boundaries may intersect or one boundary may intersect the other interior. The interiors do not touch. Undefined for point/point.		F T * * * * * * * F * * * T * * * * F * * T * * * * *
Crosses 	The interiors intersect and the base's interior intersects the candidate's exterior. Or in the case of line/line, the intersection of the interiors forms a point. Undefined for point/point or area/area. Undefined for aggregate/multi geometries.		T * T * * * * * * For two lines: 0 * * * * * * * *
Overlaps 	The interiors intersect, but neither feature is contained by the other, nor are features equal. Undefined for point/line, point/area, or line/area. Undefined for aggregate/multi geometries.		T * T * * * T * * For two lines: 1 * T * * * T * *
Contains 	The interiors intersect and no part of the candidate's interior or boundary intersects the base's exterior. It is possible for the boundaries to intersect. Inverse of WITHIN.		T * * * * * F F *
Within 	The interiors intersect and no part of the base's interior or boundary intersects the candidate's exterior. It is possible for the boundaries to intersect. Inverse of CONTAINS.		T * F * * F * * *

Figure 5: The categorized spatial relations with their named predicates and descriptions, based on the DE9IM. Bases are labelled with "A" and candidates with "B". Examples showing multiple polygons, points or lines do not represent multi or aggregate geometry, rather, they indicate alternate scenarios that match the predicate. The pattern matrix corresponds to the DE9IM matrix. In the case of "touches", the Table shows three different DE9IM matrices because all those are possible (FME, n.d.)

More recent research has focused on the understanding of these standardized terms across individuals and languages, and shows that some of these terms are not universally understood (e.g. in English, the terms “intersects” and “crosses” show ambiguous interpretation among non-experts) so can be a source of ambiguities that are not present in the 9IM by itself (Stock & Cialone, 2011). Therefore, a universal understanding of the standardized predicates can only be assumed among people that had been taught about the mathematical meaning of these terms. While geospatial professionals are assumed to know this, this is not the case for non-expert users.

To address this issue, some authors have proposed the use of diagrams to ease the understanding of the formalized relations for spatial constraints definition interfaces for non-experts (Cockcroft, 2004; Ubeda & Servigne, 1996; Servigne, et al., 2000), but some drawbacks had been identified to this approach as well, as it tends to over specify the relations (Riedeman, 2004), as *Figure 6* explains with examples. Moreover, these interfaces do not consider the general theory about end-user requirements definition which explains it as an iterative process where users refine their requirements according to new information or experiences generated by the process to fulfil the requirements itself (Preece, et al., 2015).

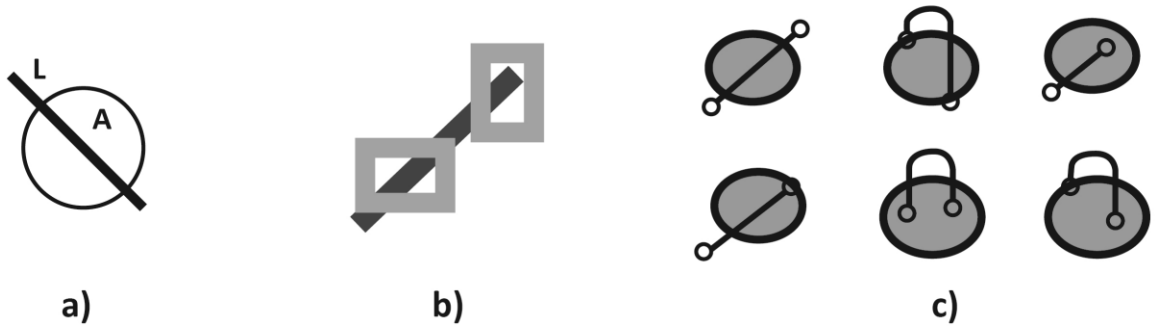


Figure 6: Examples of diagrams used by different authors to ease the understanding of the relation “crosses” between a line and an area. Figure 6a shows the diagram that Servigne et al. (cited in Cockcroft, 2004) used in an interface designed for non-experts. Figure 6b corresponds to the diagram used by FME (n.d.) *which combines two cases in one diagram. Figure 6c shows the diagrams used by Mark and Egenhofer* (1994)*. As it can be seen, Servigne and FME chose the diagram(s) that considered most representative but left outside other cases. Note that while Mark and Egenhofer are exhaustive in its description from the perspective of the 9IM, they do not pretend to be exhaustive regarding all the possible specifications and more specific cases can be imagined.*

3.3.2. Structural formalization for machine-readable processes

The base of machine-readable spatial constraints is its structure as predefined templates to organise all their elements in a systematic way. As Valleries et al. (2006) explain, apart from the topological relationships and the geographic objects that are being related, other components should be specified when defining spatial constraints. The first one is the cardinality of the relationship which specifies

the number of objects (if any) with which an object can have the specified relationship and is represented by a pair of values; the minimum and maximum number of objects with which an object can have a relationship (Vallieres, et al., 2006). The maximum number can be replaced by "none (n)" to define that there is no maximum, and, as extended by Bogorny et al. (2005) both can be replaced by "all (a)" for the case of the relation disjoint.

Additionally, a buffer can be defined to deal with coordinate's precision problems or as a level of tolerance. As the authors explain, the different software uses different levels of precision (relevant decimal places) which can generate problems when using spatial operators as the underpinning algorithm treat as equal only the coordinates that are identical, including all their decimal figures (Vallieres, et al., 2006). *Figure 7* shows the structure reviewed by Valleries et al. (2006).

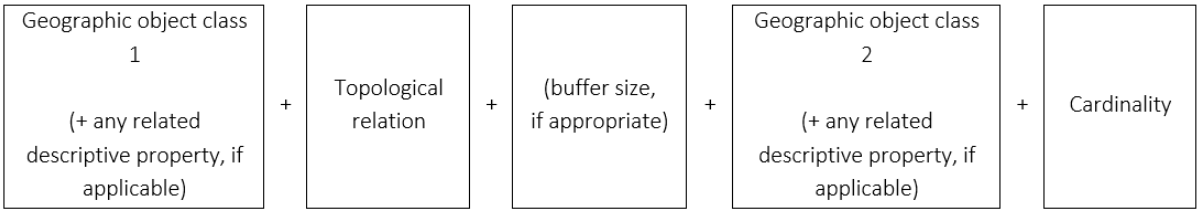


Figure 7: Formal structure to document spatial constraint. Based on Valleries et al. (2006).

An in-depth review did not identify any papers related with the communicational value or limitations of this structured format, probably because it is not the purpose of them, as it implicitly is when documenting spatial constraints as metadata.

3.4. Documenting Data Quality - Metadata

Metadata is additional information describing a data resource that helps a potential user to 1) discover it using a searching engine, 2) evaluate if it fits his/her purposes, and 3) use it properly (AGI, 2012, p. 6). As can be assumed directly from this definition, metadata is created by the producers of a dataset for the (potential) users.

As noted in *Chapter 1*, the goal of INSPIRE (to which the LBH and all government entities in UK must comply) is to facilitate the access and reuse of spatial data in Europe so it tries to address the issues mentioned above. As part of INSPIRE, any member entity shall report DQ information as metadata in a standardized way (to allow cross-comparison of datasets) and handled separately to the data itself to allow the assessment of its fitness for the user's purpose before having the data (JRC European Commission, 2013).

In the context of lack of shared terminology for DQ documentation (JRC European Comission, 2013), INSPIRE provides its community with the guidelines as to how to document DQ as metadata, and a common assessment methodology among all the participant entities, to ease the comparison between datasets for users. These guidelines were specified in ISO 19157: 2013 and to ensure convenience comparability, only one measure is defined to document (within the metadata) the compliance of a dataset with each DQ element (BSI, 2018).

3.4.1. Metadata and Topological Constraints

Given the focus on topological relationships, the ISO 19157 DQ element relevant for this project is the *logical consistency,* defined as the "degree of adherence to logical rules of data structure, attribution and relationships" (BSI, 2018). Spatial integrity constraints are part of these logical consistency rules and define, through axioms, what would be logically consistent for a particular model (e.g. "a building should not overlap another building", or "a streetlight should not be within a building").

Within INSPIRE, as noted above, only one measure is required as metadata for each quality element. Thus, in the case of logical consistency, it is necessary to define an overall measure that groups the compliance for all the logical rules specified by the product specifications. Among the DQ basic measures suggested by ISO, the most suitable for the assessment of rule compliance are the counting-related such as error rates or error counts. At the same time INSPIRE recommends the publication of multilevel metadata, distinguishing one level for the general public (with a more descriptive approach) and another for professional users (more detailed and formally specified) (JRC European Commission, 2013).

It is important to note that while INSPIRE requires the documentation of the degree with which a dataset conforms in general to the associated topological constraints, it does not require the documentation of the constraints themselves. In some situations –e.g. Ordnance Survey (2017)– this is done within a data specification document. However, within the LBH the aim is to extend the standards-based metadata to include this additional information, as data specification documents do not exist due to the lack of experts and data governance concerns (see *Chapter 2*).

3.5. Summary of the literature review

To summarise, topological constraints are a key tool to improve DQ and its conformity should be informed via metadata according to INSPIRE. But their expression for the purpose of automatic processes is not easy and the existing approaches to help non-experts in its definition are not exempt of limitations. Moreover, when these approaches have stored the constraints as metadata this has been done due to practicality but not giving special attention to its communicational value as

metadata. For those reasons, the approach that seems more suitable instead is the expression of constraints in natural language by the LBH's DOs, considering an iterative process to make clear possible ambiguities. This approach will also allow obtaining learnings on how to facilitate the communication of constraints to other users (which has not been studied) to be considered as part of the proposal for a process to assist the automation of spatial DQ rules validation based on metadata about spatial constraints, developed in the following chapters.

Chapter 4: Methodology

This research was based on a qualitative methodology applied to pilot DOs from different council services. The process was divided into three phases: 1) expression of spatial rules by the DOs, 2) formalization of them, and 3) definition of quality checking routines and reports. As explained in *Chapter 3*, even though there is literature about interfaces to help the definition of constraints directly in a formalized way, in this project it was decided to separate both steps as the formalized terminology to express constraints have shown not been clear enough for non-experts (see *Section 3.3.1*) and as a way to learn communicational strength from the manner they naturally express this kind of rules.

For the first and last phases, the methodology implied gathering data from pilot DOs to define their requirements for the data and to receive feedback about the report's proposals. In the first phase five DOs from different LBH services participated, all the expressed rules were then formalized in phase 2, but only subset of them was checked and reported due to time limitations.

All the gathering data process was agreed with the participants through an informed consent form which assures their anonymity (see *Appendix 1*). The sample was composed by voluntary DOs, but it was ensured that they had a diversity of profiles as the general situation in the LBH, even though this sample does not pretend representativeness (see *Appendix 2* for the questionnaire applied to the participants to illustrate their level of expertise/experience in GIS and a summary of the results).

Considering that the definition of user requirements is an iterative process, the expression of DO's requirements (phase 1) considers three steps: a preparation questionnaire, a workshop, and additional individual interviews, if considered necessary. This methodology was designed not only for the purpose of this research but as a test for a future replicable methodology to express constraints to be used in the future by the LBH GIS corporate team. For this reason, the information gathered was not only regarding the participant's answers but also to the observation of their reactions to the methodology.

The preparation questionnaire required the DOs to gather preliminary information from their teams about what are the team needs regarding the data, to be used in the workshop. This questionnaire presented the definition of quality used by this project and asks two questions to the DOs: 1) What are your team requirements related to operational needs? and 2) What are your team requirements related to general quality considerations in order to make the data clean, easy to understand, publishable, or used to feed other systems without crashing? For each case, some examples were given avoiding its formulation as spatial rules but as hypothetical situations about what they would like from their datasets. See preparation questionnaire in *Appendix 3*.

Two workshops were held then with the participation of two and three DOs respectively. These workshops represent a combination between different research methods: a semi-structured group interview with other ways to express the answers (written and with drawings) along with the use of observation of the reactions of the participants to the activity and supportive material. This method was preferred over personal interviews to foster discussion and to stimulate the less participative or with less expertise/experience in the area with the examples given by the rest, so the groups were made ensuring different levels of expertise/experience.

In the workshops, the main activity asks the participants to identify and express the rules that were behind their idea of a good enough dataset for their purposes, with the following instructions:

"Imagine that you need to explain to a new data custodian (with limited experience in GIS) how the information should be represented in the datasets. Your purpose is to make sure that the relevant objects are correctly represented in the map reducing the most possible the editing errors, especially in the features that are more important for the department's needs. We ask you to define this as a list of "rules", writing and drawing them in a form (one rule per form). Write in the back any explanation or comment you think is important and draw it in the way you think is the clearest to understand the rule. In the end, we will share what all of us did." (See *Appendix 4* for the full outline, *Appendix 5* for the supporting presentation, and *Appendix 6* for the form used by each participant to register his/her rules).

To facilitate the task, the DOs were provided with laptops with access to their datasets on Earthlight.

After the workshops and the first attempts to formalize the expressed rules and check them, individual interviews were held to solve questions emerged in the formalization and check phase.

The formalization of constraints (phase 2) was done using the named predicates defined by ISO 19125 and the structure suggested by literature (see *Figure 7* on page 13). The checks (phase 3) were then designed using FME Desktop. A basic check was done to five of the expressed rules trying to cover the whole range of them in terms of the geometries involved, relations and the inclusion of "where clauses" (filters to extract only the records that fulfil a specific condition).

Given the number of issues found for some of the rules was high, a second version of more detailed checks was done to distinguish different kind of errors or severity levels, to see if this can improve the usefulness of the reports with the hypothesis that distinguishing types of issues can guide the decisions on how to solve them, to which ones give more priority or even define different ways to solve them depending on the type.

For the final phase, two different report prototypes were manually created for three of the pilot services as the automation itself is out of the scope of this project and a semi-structured individual interview was held with each DO (see interview script in *Appendix 7*) to obtain their opinions on how easy to understand and useful these reports are as a tool to improve DQ.

Chapter 5: Results

This chapter presents the results organized in three phases: 1) expression of spatial rules by the DOs, 2) formalization of them, and 3) definition of quality check routines and reports (basic and detailed).

5.1. Expression of spatial rules by the data owners in natural language

5.1.1. Evaluation of the method used to guide the expression of spatial rules

The method used to guide the expression of constraints (preparation questionnaire to define requirements, followed by a workshop to define constraints based on these requirements, and individual interviews for clarifications) showed diverse responses between the DOs: only some filled the preparation questionnaire (see answers in *Appendix 8*) and among those who did it, it was filled more in-depth and asking their teams by the DOs with less time in their positions (independently of their level of expertise in GIS), however, they seemed confused by it, asking for clarification before the workshops and presenting more difficulties to express their constraints.

The workshops itself presented good results in terms of the constraints expressed but they did not show a special utility fostering the less experienced by the examples given by the others or generating proper discussions. Except for some moments, the conversation flowed more like an individual interview in turns and there were not many topics risen by one or some of the participants and considered later by others.

Regarding the use of their laptops to check their data on Earthlight, some of them used it actively while others did not, without any clear relationship with their expertise/experience level or time in their positions.

5.1.2. Expressed constraints

Among the expressed constraints, 11 out of 19 correspond to binary topological relations and 3 are non-binary relations that can be binarised as they refer to the relation of one feature class with the union of all the features of another class. For a matter of space, all these rules are presented in *Table 2,* in next section, along with its formalization. The non-binary or non-topological relations expressed can be found in *Appendix 9*. See in *Appendix 10* all the forms as were filled by the DOs.

A first interesting aspect is that, opposed to its formal definition, DOs understand constraints in a less strict way, claiming the existence of exceptions that require to be handled, as shown in the citations below.

"DO2: Highway assets should only appear within the public highway polygon, most of the time. There are some exceptions but most of the time it is like that.

GIS Team member: But, how do you define those exceptions? Is there any special kind of asset that does not need to follow the rule? […]

DO2: No. I just know where the exceptions are. I cannot really think of a rule for the exceptions."

Workshop n°1. (Transcription in *Appendix 11*)

"My concern is if you are going to create GIS rules to prevent data entry based on these rules because there are some situations that have exceptions. If the system becomes too rigid, it is complicated. I worked previously in a planning database in another council where were rules to stop certain data entries and you ended with the most load work trying to cheat the system, using redundant fields to explain things, and it was really scrappy and painful."

DO n°5, Workshop n°2. (Transcription in *Appendix 12*)

"[…] a recycling bin should be within it [a recycling estate], though there are some exceptions. It would be helpful if we can define exceptions to the query, so they do not come up again in the report."

DO n°3, Workshop n°2. (Transcription in *Appendix 12*)

Regarding the terms used, DOs not always used the standardized predicates to communicate the rules. The only one they used for binary topological relations was "within" (and sometimes they refer to it as simply "in") even though there were rules implicitly referring to other relations as it is going to be seen in the next section. Furthermore, from the use they give to the term "within", it seems the term itself is not enough to communicate the well-defined mathematical meaning of it (see *Figure 5*, page 11) when referring to polygonal or linear features, requiring a clarification, as one of the DOs explains:

"Interviewer: about this rule "LBH managed roads should be wholely within LBH public highway" why you think is important to say "wholely within" and not just "within"?

DO2: just to be clear, and because it can be understood that when a part is inside it is also within."

Workshop n°1. (Transcription in *Appendix 11*)

Regarding the diagrams used to express the rules, not all the DOs used this resource limiting what can be extracted from them. However, among those who used it, it is remarkable the intent to resemble the real features' shapes instead of abstract geometries, the addition of contextual features, and the inclusion of both the compliant and not-compliant examples with signs as checks and crosses to indicate them (see *Figure 8*).

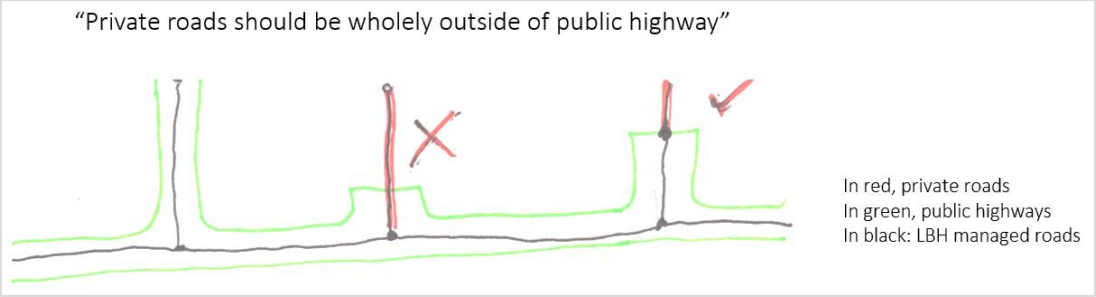


Figure 8: An example of the supportive drawing for one of the rules expressed

5.2. Formalization of spatial constraints

The formalization of rules presented in *Table 2* was based on the predicates defined by ISO (see *Section 3.3.1*) and the predefined structure reviewed by Valleries et al. (2006), as follows:

"Feature Class 1 + (where clause) + ***topological relation*** + (buffer) + Feature Class 2 + (where clause) + [cardinality]"

Figure 9: Structure used to formalize the rules

As it can be derived from *Table 2*, the rules expressed by the DOs imply only part of the topological relations possible according to the 9IM. There were no constraints between points, between lines, or between points and lines, and there were no constraints related with the relations "disjoint", "touches" and "contains" (see *Appendix 13* for the coverage of topological relations by the expressed rules). On the other hand, no buffers were explicitly defined by the data owners to handle a certain level of tolerance or coordinate inaccuracies.

Table 2: LBH's rules and their formalization

N°	Natural language expression	Formalized expression
BINARY TOPOLOGICAL RELATIONS		
1.	A recycling bin should be within a recycling state[1].	**Recycling bin** *within* **Recycling estate** [1,1]
2.	Addresses with BLPU different to RD06 (flats) should not be within recycling states[2].	**Address** (where BLPU ≠ RD06) *within* **Recycling estate** [0,0]
3.	Polygons identifying the demise of a shop should be within a building	**Building** (where use = shop) *within* **Corporate building** [1,1]
4.	Polygons identifying a building alone and the ownership boundary with no extension into the road	a) **Corporate buildings** *overlap* **OSMM topographic area** (where type = road) [0,0]
		b) **Ownership boundary** *overlap* **OSMM topographic area** (where type = road) [0,0]
5.	Car parks areas should not include any topographic area covered with grass.	**Car parks areas** *overlap, contain, within or equal* **OSMM topographic area** (where type = grass) [0,0]
6.	A conservation area must be in Hackney	**Conservation area** *within* **Hackney borough boundary** [1,1]
7.	A building should not be in more than one conservation area	**OSMM topographic area** (where type = building) *within* **Conservation area** [0,1]
8.	A conservation area boundary should not cut a building.[3]	**Conservation area** *overlap* **OSMM topographic area** (where type = building) [0,0]
9.	Hackney's data on nationally listed buildings does not exactly match Historic England data (and it should)	**Nationally listed building** *equal* **England historic building** [1,1]
10.	The shape of a listed building should be the shape on OSMM (with exceptions)	a) **Nationally listed building** *equal* **OSMM topographic area** (where type = building) [1,1]
		b) **Locally listed building** *equal* **OSMM topographic area** (where type = building) [1,1]
11.	Private roads should be wholely outside of public highways.	**Private road** *cross* or *within* **Public highways** [0,0]
12.	Point data (general practitioners / opticians / pharmacies / hospitals / dentists / clinics) [should be] within borough boundaries.	a) **General practitioners** *within* **Hackney borough boundary** [1,1]
		b) **Opticians** *within* **Hackney borough boundary** [1,1]
		c) **Pharmacies** … (one rule for each feature type)
NON-BINARY TOPOLOGICAL RELATIONS THAT CAN BE BINARISED		
13.	Highway assets should only appear within the polygon for public highway	a) **Highway asset type 1** *within* (the union of) **Public highways** [1,1]
		b) **Highway asset type 2** … (one rule per asset type)
14.	LBH managed road wholely within LBH public highway	**LBH managed road** *within* (the union of) **Public highways** [1,1]
15.	All housing estate assets should be within housing estates polygons	a) **Asset type 1** *within* (the union of) **Public highways** [1,1]
		b) **Asset type 2** … (one rule per asset type)

[1] This rule was originally expressed as "Identify all recycling bins not within estates or not within 5m of [recycling] estates" which instead of a rule corresponds to the analysis to be done to check the implicit rule.
[2] This rule was originally expressed as "Identify recycling estates not RD06 (flats) which instead of a rule corresponds to the analysis to be done to check the implicit rule.
[3] This rule was originally expressed as "A conservation area boundary should not cut a building in half" but latter it was clarified that it was not necessary to be in halves.
Note: text in square brackets corresponds to words not present in the original DOs' expression but added to facilitate the understanding.

It is also noticeable that in some cases the words chosen by the DOs are more concise than when limiting to the standard predicates. For example, “private roads should be wholely outside of public highways” express more concisely that they should not cross or be within public highways and probably in a more clear way if it is considered the ambiguous interpretation of the term “cross” described by Stock and Cialone (2011) (see *Section 3.3.1*) and the ambiguity of the term “within”, as described in the previous section. As another example, in “car parks areas should not include any topographic area covered by grass”, the term “include” sum up something that, even though is mathematically simple (the intersection of the features’ interiors should be empty), with the standard predicates should be: “overlap, contain, be within or be equal to”.

5.2.1. Difficulties identified in the process of translation

a) As experimented during the translation process, even the standardized predicates are mathematically well defined, it is difficult to avoid confusion with the everyday use of those words. For this reason, when formalizing the rules, a method was followed (*Figure 10*) consisting in the systematic check of all the geometrically possible relations for the type of features under analysis to define if they are allowed, compulsory or forbidden and its cardinality. The relation “intersects”, was checked at the end of the process as it is actually defined as the presence of any other relation except form disjoint, so it needs the definition of the rest to be defined (see in *Appendix 14* a table with the possible relations categorized as forbidden, allowed or compulsory for all the rules expressed). The previous process was effective ensuring that the rule expressed was completely translated with all the spatial relations that it implies.

b) As expressed by the DOs, rules not always explicit cardinality thus clarification with the DOs was required (e.g. a recycling bin should be within a recycling estate” which then was clarified as “… within one recycling estate”)

c) The same for the “where clauses” when referring to the OSMM topographic layer as a referential layer. DOs expressed the rules referring to things in the real world assuming that these things can be identified in the OSMM layer. However, this was not always possible (e.g. it is not possible to identify grass areas to check the rule n°5)

d) The predefined structure is not designed to express relations between an object class and the union of another, as required to express the non-binary topological relations that can be binarised through this operation. For this reason, in this case, the feature class 2 were expressed as “the union of + feature class 2”.

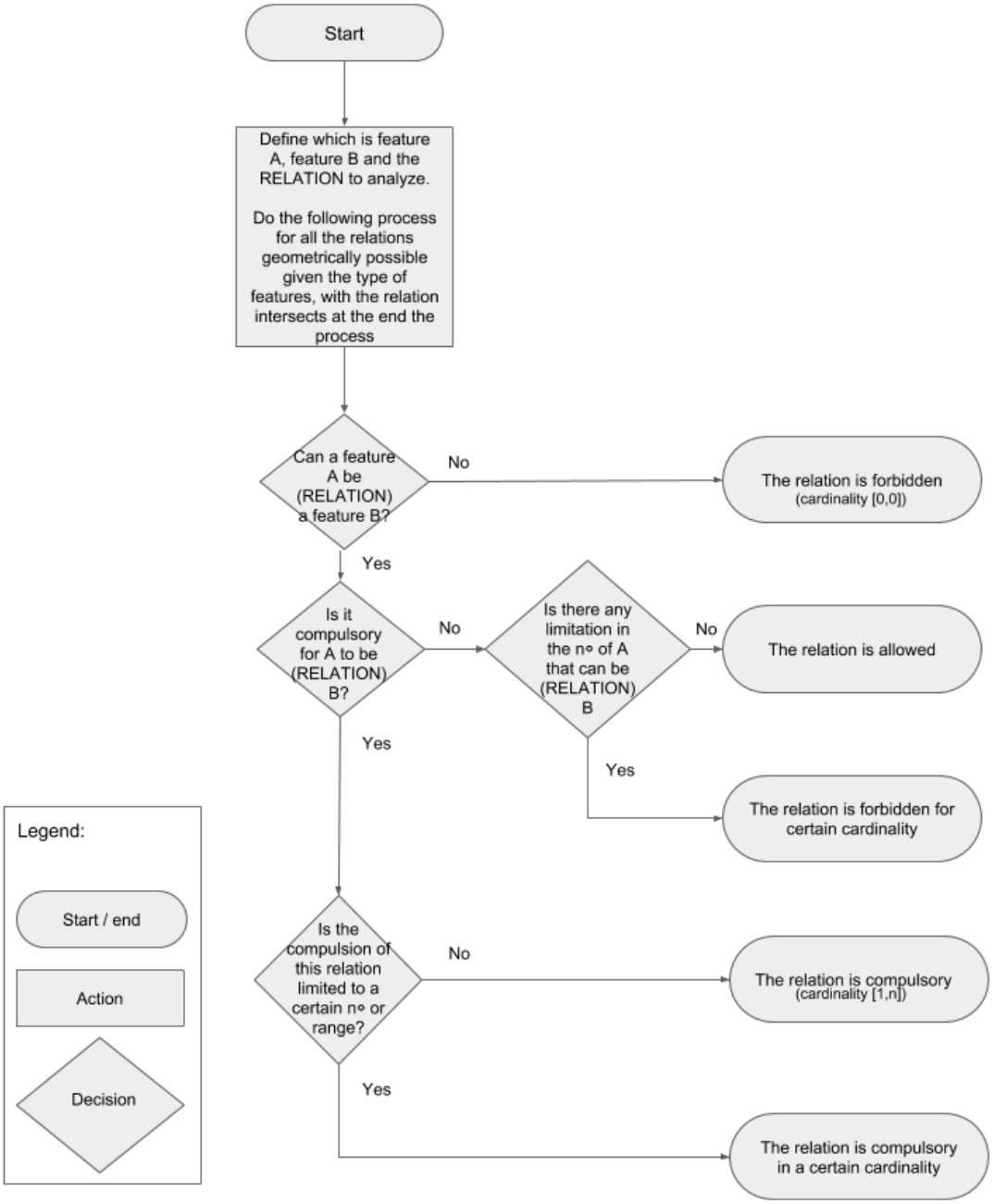


Figure 10: Proposed method to ease the translation of spatial relations from natural English to the standardized terminology. E.g. To check the relation "disjoint" according to the rule "a recycling bin should be within a recycling estate" the logic would be: "Can a recycling bin be disjoint a recycling estate" → yes (it should be within one but that does not mean it cannot be disjoint to all the rest). Is it compulsory for a recycling bin to be disjoint a recycling estate? → No. Is there any limitation in the number of recycling bins that can be disjoint a recycling estate? → No → so the relation is allowed

5.3. Quality checking routines and reports

5.3.1. Basic checks and reports

In FME, the process to check the compliance of the rules could be organized in 4 steps: a) *data reading*, b) *data preparation*, c) *topological relation test,* and d) *data writing* (see the FME workspaces of the

three rules reported in *Figure 11*. The rest of the basic checks workspaces can be found in *Appendix 15*).

The data reading process included the definition of filters to read a sample of the data. These filters can be defined in two forms: a) as "where clauses" or b) as a spatial filter useful for external data sources with information beyond the LBH, or in case some rules apply only to some areas.

The data preparation process included a) the disaggregation of aggregated features (because a topological relation test requires single features) and/or b) the union of features with the dissolver tool when dealing with non-binary relations that can be binarised through this process (e.g. Highway assets should only appear within the union of public highways).

The topological relation test included the use of a *spatial relator* and a *tester*. The spatial relator tool identifies, for each feature A (called requestor), if it presents one or more of the standardized relations (determined by the user) with any feature B (called supplier). The output of it are all the requestor features with a new attribute called "number of related suppliers" informing the number of features B (suppliers) with which each feature A (requestor) have this relation. Then, it is in the tester tool where the number of related suppliers that make a feature pass the test is defined (e.g. for the rule checked in *Figure 11a*, the test was defined to be passed if the number of estates in which the bin was =1).

Finally, the data writing consisted of saving the non-compliant features (along with its key information) adding a unique identifier through the counter tool.

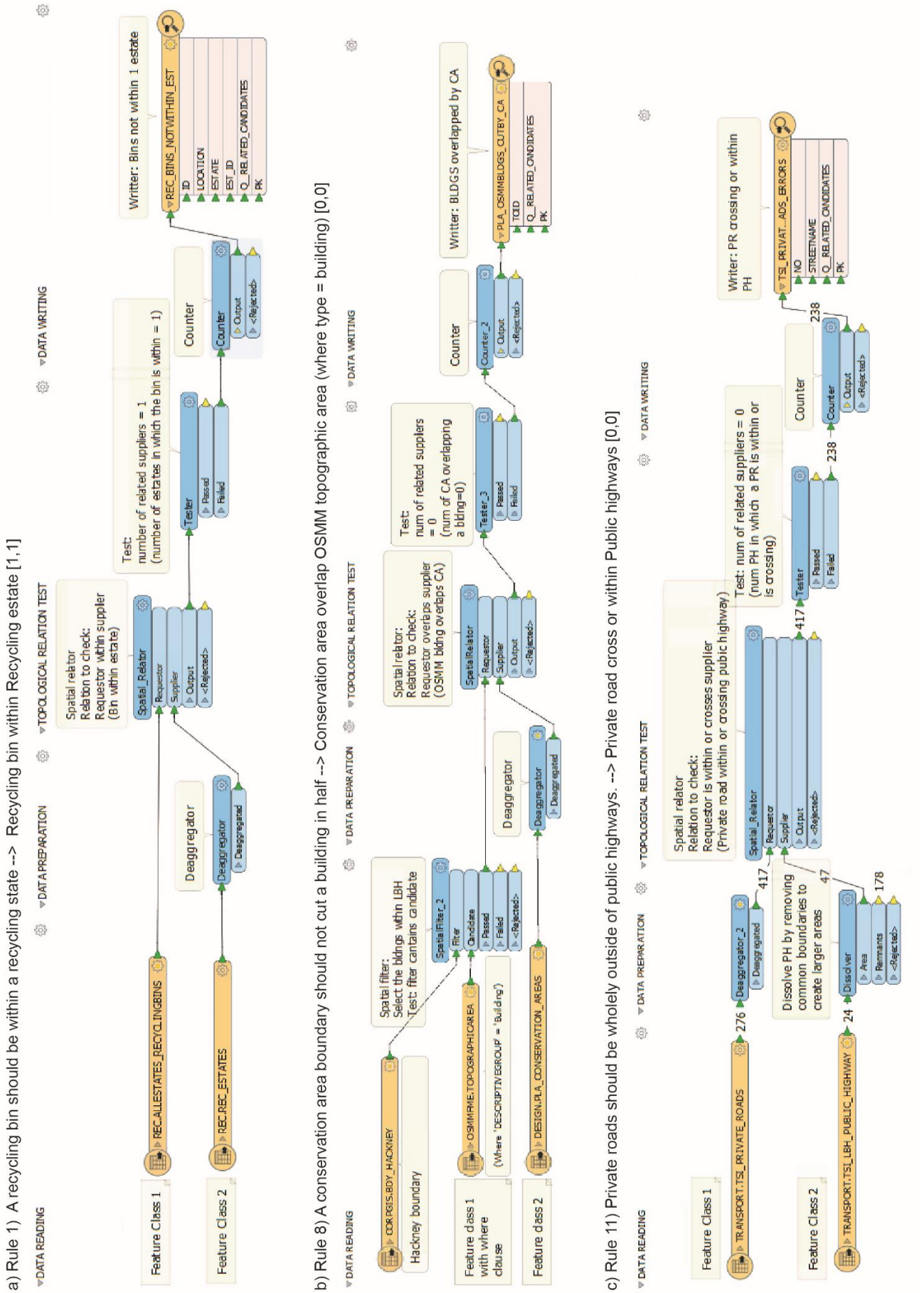


Figure 11: Workspaces for the basic checks of three of the rules. All of them follow the same structure.

Considering this process, the structure used to formalize the constraints presented in *Figure 9* (page 21) is suitable to be automatically read by FME as explained in *Figure 12*. The only change that may ease this process is the expression of the cardinality in the same way FME asks for it in the test tool. In it, the test defined should be the number of related suppliers that make a specific feature to pass the rule, and this test should be expressed using the signs =, ≠, <, >, ≤, ≥, and the connectors "OR" or "AND".

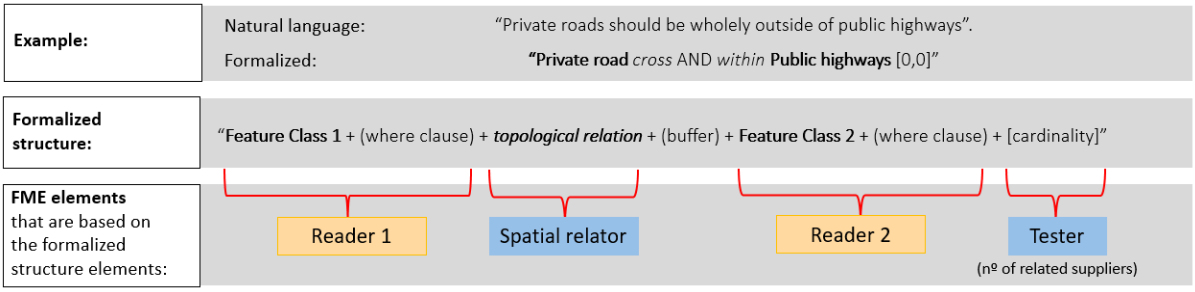


Figure 12: Correspondence between the formalized structure and the key FME workspace elements.

Additionally, the check of the rule "**Conservation area** *overlap* **OSMM topographic area** (where type = building)[0,0]" showed the importance of defining clearly which feature class corresponds to the requestor and which to the supplier as in this case it was more informative to identify buildings that are overlapping the conservation areas than the conservations areas that are overlapping buildings because the latter analysis does not show the specific problems, even if it is the conservation areas which should be corrected.

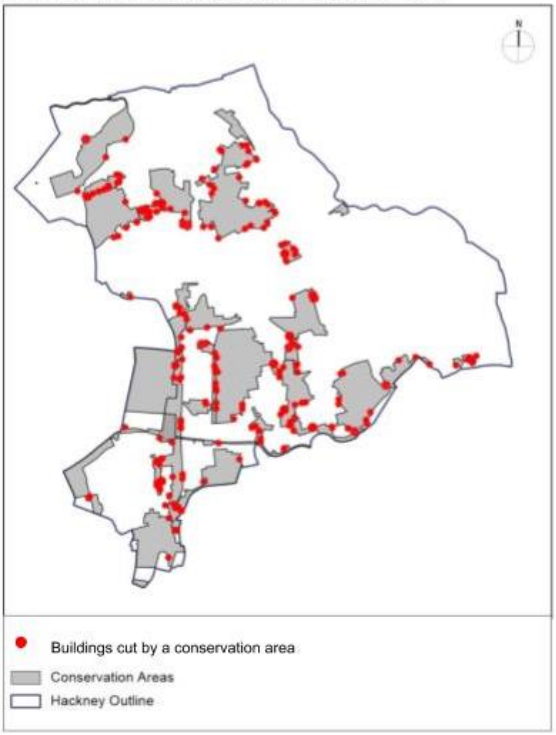
The results of the five rules checked in terms of number and percentage of non-compliant features is presented in *Table 3*, and *Figure 13,* in the next page, shows one of the basic reports proposed. The rest of them can be found in *Appendix 16*.

Table 3: FME basic checks results

Formalized rule (with feature class 1 = requestor)	**N° of requestors failing the test**	**% of requestors failing the test**
1. Recycling bin *within* **Recycling estate** [1,1]	330	17,3%
8. OSMM topographic area (where type = bldng) *overlap* **Conservation area** [0,0]	245	0.4%
10b. Locally listed building *equal* **OSMM topographic area** (where type = bldng) [1,1]	1224	100%
11. Private road *cross* or *within* **Public highways** [0,0]	238	46,3%
14. LBH managed road *within* (the union of) **Public highways** [1,1]	367	40%

Data quality checking report

"A conservation area should not cut a building"



The conservation areas are cutting

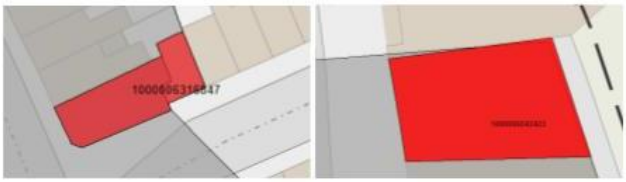
245

buildings in Hackney.

Where to find the data: *(link to location or to download)*

Examples:

TOID	Conservation area	CA_ID
5000005170592835	Clapton Square	1
5000005186494818	Graham Road and Mapledene	18



TOID	Conservation area	CA_ID
1000006316847	Clapton Square	1
1000006042423	Dalston	31

TOID	Conservation area	CA_ID
1000006042402	Dalston	31
1000001802845624	Dalston lane (west)	24

Alternatives to solve these issues:

- Manually correct the conservation area polygon to eliminate overlaps.
- *Mark the building as an exception* (link)

Figure 13: Basic report

5.3.2. Detailed checks and reports

As shown in the previous section, the results of the basic checks showed high figures of non-compliant features and a closer analysis of them identified different types of issues. Some of the issues were due to different ways to handle coordinates decimal figures (as reviewed in *Section 3.3.2*) between different sources. This was particularly evident and conflictive when checking rule 10 related with the equality between OSMM buildings and conservation listed buildings which gave zero compliant features while with a visual check it was difficult to find differences in most of the buildings and finally were millimetre displacements.

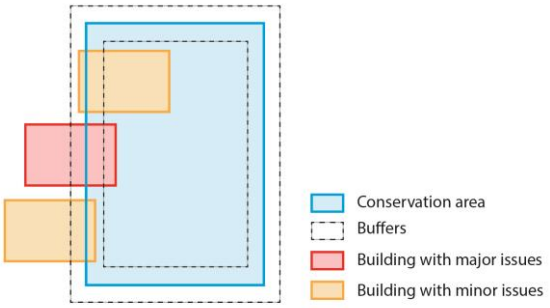
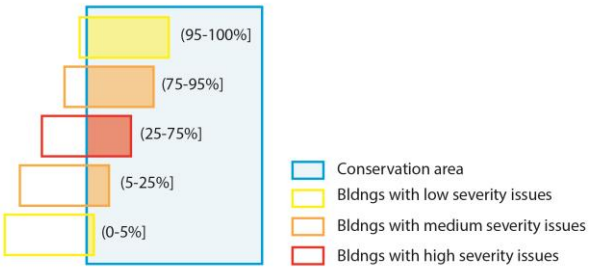
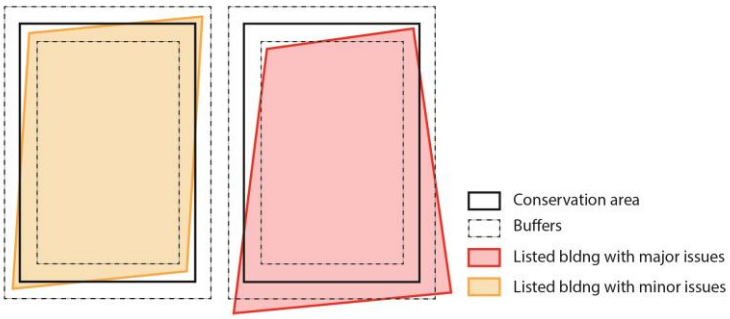
Other issues seem to be because of the manual definition of features without a proper snapping tool, as for example in the case of rule 8 the conservation areas overlapped an important proportion of OSMM buildings in very small areas.

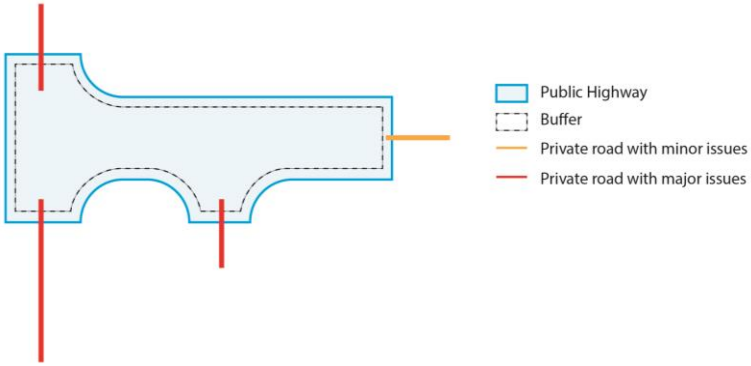
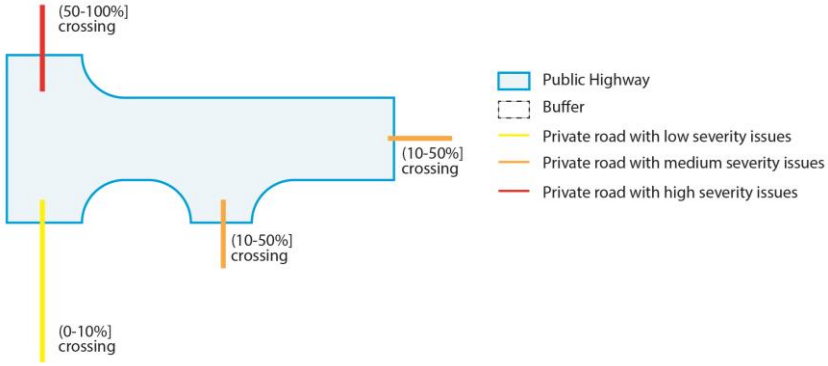
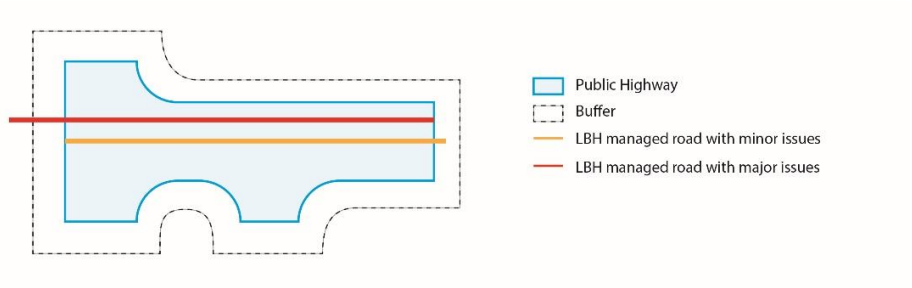
New checks based on buffers as considered in the formalized structure reviewed by Vallieres et al. (2006) were developed to categorize the issues. Alternative percentage approaches were tested for some of the checks as a way to define severity levels. At the same time, when a rule expressed more than one topological relation, the violation of each of them was reported separately.

As a general result of these tests, the use of buffers resulted more straightforward than the percentage approach and simpler in terms of quantity of transformers used. Additionally, while useful to define "severity levels", the percentage approach appeared less suitable for dealing with coordinate precision problems and the lack of snapping tools when defining features, as it is affected by the size of the features.

The following pages present, first, a table summarizing what was done in each case (*Table 4*), followed by an example of a workspace to create a detailed report based on buffers along with one based on a percentage approach for the same rule in *Figure 14* (page 32). Finally, the detailed report produced with this detailed check, specifically with the percentage approach is presented (*Figure 15*, page 33). All the rest workspaces with the detailed parameters used can be found in *Appendix 17*. The rest of the detailed reports can be found in *Appendix 18*.

Table 4: Summary of the processes followed to identify different types of issues

Formalized rule (with feature class 1 = requestor)	**Summary of the process to identify types of issues**
1. Recycling bin *within* **Recycling estate** [1,1]	There was no justification to support the need of a tolerance margin. The issues were detailed as follows: The features failing were then categorized "not within an estate" or "within more than one estate" (which resulted in only one case).
8. OSMM topographic area (where type = building) *overlap* **Conservation area** [0,0]	a) Two buffers of 0.5m (one outer and one inner) were created around each conservation area (CA). Then, the buildings overlapping both buffers of the same CA were identified. This retrieves only the major issues which resulted been 22% of the total issues. The minor issues were derived subtracting these issues from the total issues identified with the basic check.  b) The percentage that the overlapped part represents from the total building was calculated. Then, these percentages were categorized in a new attribute called SEVERITY_LEVEL distinguishing three levels –low, medium and high– which respectively resulted been 93.5%, 2.9% and 3.7% of the total issues. As a building can overlap a CA in more than one part, three different datasets where created for each severity level so a building can be in more than one. 
10b. Locally listed building *equal* **OSMM topographic area** (where type = bldng) [1,1]	Two buffers of 1cm (one outer and one inner) were created around each OSMM building as a "tolerance margin". Listed buildings not within this tolerance margin (not within outer buffer or not containing inner buffer) were identified and considered major issues. They represented 14.2% of the total issues. The minor issues were derived subtracting these issues from the total issues identified with the basic check. 

11. Private road *cross* or *within* (the union of) **Public highways** [0,0]	a) A 1m inner buffer was created inside the public highway (PH) polygons (previously dissolved). Then, the private roads (PR) crossing the inner buffer were identified and considered major issued, representing 92.4% of the total issues. The minor issues were derived subtracting these issues from the total issues identified with the basic check.  b) The relations cross or within were tested separately. The PR crossing the PH were categorized by severity level according to the percentage of the PR inside the PH. 64.3% was low severity, 29.3% medium and 6.4% high. On the other hand, the PR within PH were saved apart.  c) The last approach resulted not useful in this case as PRs crossing the PH in similar extents were categorized differently because the percentage depends on the full extension of the road. A final version was done based on the PR length inside PH resulting in 85.1% low severity issues, 5.2% medium and 9.6% high, among the PR crossing PH.
14. LBH managed road *within* (the union of) **Public highways** [1,1]	An outer 15m buffer was created around the PH (previously dissolved). Then, the managed roads not within the PH's outer buffer were identified and considered major issues (representing 36.2% of the total issues). The rest of the issues were derived subtracting these issues from the total issues identified with the basic check. 

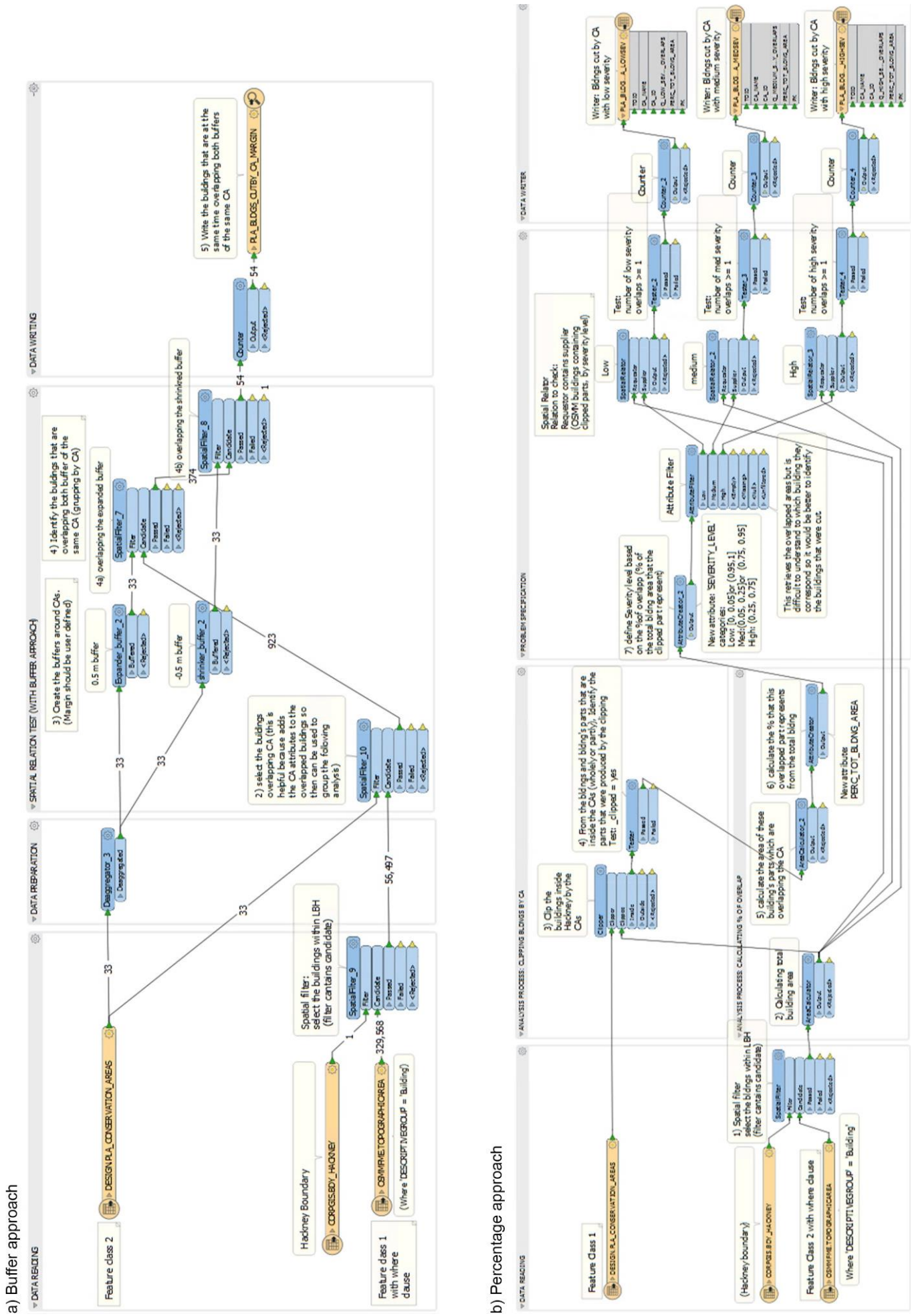
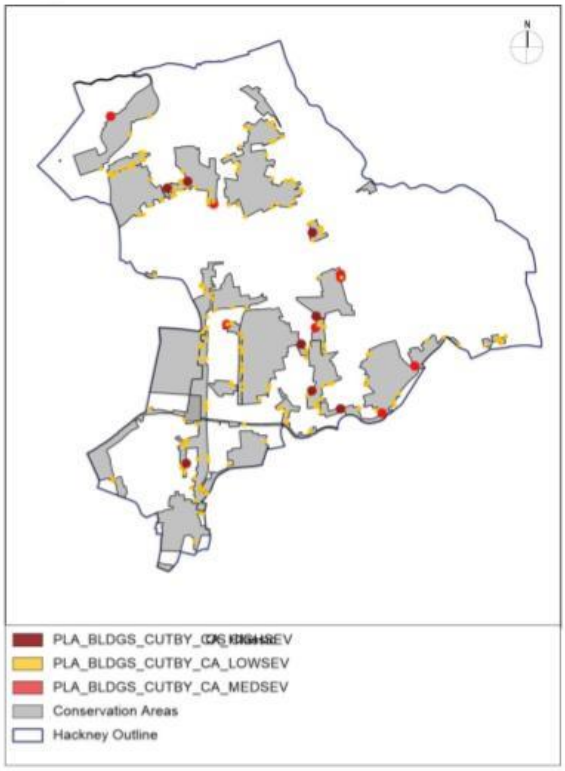


Figure 14: Detailed checks workspaces for the rule 8) "a conservation area should not cut a building".

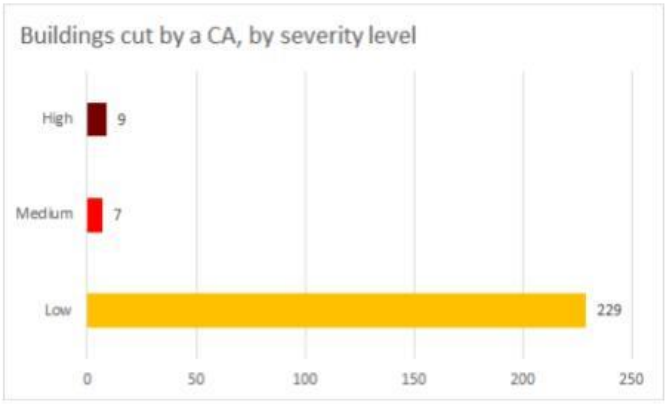
Data quality checking report

"A conservation area should not cut a building"



The conservation areas are cutting 245 buildings in Hackney, 9 of them with high severity, 7 with medium severity and the rest with low severity (see criteria for severity level at the bottom).

Notice that a building can be cut in more than one part so a building can have issues with different severity levels.



Where to find the data: *(link to location or to download)*

Examples:

Situation 1: High severity level
(25% - 75% of the building is inside the CA)

TOID	Conservation area	CA_ID
5000005170592835	Clapton Square	1
5000005186494818	Graham Road and Mapledene	18

Situation 2: Medium severity level
(5% - 25% or 75% - 95% of the building is inside the CA)

TOID	Conservation area	CA_ID
1000006316847	Clapton Square	1
5000005186572816	Victoria Park	6

Situation 3: Low severity level
(Less than 5% or more than 95% is inside the CA)

TOID	Conservation area	CA_ID
5000005151934514	Stoke Newington Reservoirs, Filter Beds and New River	11
1000006042423	Dalston	31

Alternatives to solve these issues:

- Manually correct the conservation area polygon to eliminate overlaps.
- *Mark the building as an exception* (link)

Figure 15: Detailed report

5.3.3. Reports' feedback

From a general view, the three DO aggreed in the usefulness of the reports (both the basic and detailed) as a tool to raise a problem, communicate it to the rest of their teams and work together to solve it, as the quote bellow express it:

> *"We [the recycling team] will work together to correct the data so they are going to be more clear on how the data should be captured. What we do not want is non-compliant records coming in, so the idea is to give the recycling team ownership on maintaining this layer. So, when they capture they would know what the constraints are."*
>
> DO3, report feedback interview (Transcription in *Appendix 20*)

For this purpose, the general map with the location of the issues and the graphs were considered relevant to understand the distribution and the scale of the problem while the examples helped them understand the issues. Related with the latter they all acknowledge the advantage of a)having the base maps and cartographic text as a way to recognise the location, b) using the same style used in earthlight because it is the way they and their teams are used to see the data, and c) plotting as well the compliant features. There was no agreement about the chart type to use.

Regarding the data provided to recognize the features with problems (through labels in the map and tables), there was also agreement in that the most important data was, more than the IDs, the information that helps them recognise the area (such as addresses, streets, postcode, estates, CA names). They also agreed in preferring the less possible information as too much information may confuse and because they are going to see the full data in Earthlight anyway.

In relation to the detailed reports, they agreed that they help to set priorities but not necesarily with the concept of *severity*. Instead, they refer more to the recognition of different types of issues as the following quote exemplify.

> *"The concept of severity does not work on us because in this case, it should be absolute, how much is in or out does not matter. We are not interested in what percentage is inside. That one is in or out? If it is both, that is a mistake. The severity does not help in this case. […] That [the cases flagged as low severity] might worth to show it because that is clearly in or out […]. But that would be a lower priority. […] I think they are two different type of issues, first, when the building is cut in half, or in pieces, and second, when the boundary line does not seem to be following the land registry boundary. So I would put together the high and medium severity as the first type of error, and the low severity just as another type of issue".*
>
> DO5, report feedback interview (Transcription in *Appendix 21*)

Chapter 6: Discussion

First of all, the fact that all DOs expressed the existence of exceptions for their rules questions the traditional definition of constraints in RDBMS and adds an advantage to the approach based on DQ checks routines over the direct implementation of the rules in the database to prevent uncompliant data entries, as the DOs agreed.

To facilitate the process of automated quality check routines based on constraints documented as metadata –which essential communicational purpose should not be forgotten–, what the results suggest can be organized as follows:

1. Regarding the methodology used to help the identification of spatial constraints by the DOs, the results showed that it was easier for people with more time working with the data, and, in particular its first step (preparation questionnaire) was even confusing for the ones with less time working with the data. The workshops worked well but not especially for the purpose of fostering the participation of the less participative and revealed that the assumption that it was going to be more difficult for the less experts in GIS was wrong. The findings indicate that the relevant experience to identify spatial rules is the use of it, more than general knowledge in GISc.

2. Regarding the way the constraints should be expressed in the metadata, the results confirmed that the structure to formalize constraints reviewed by Valleries et al. (2006) and the standardized terminology based on the DE9IM and suggested by ISO allow to define all the fields required by FME to check the compliance of the rules analysed. Thus, it was possible to define an FME workspace template useful for general automatable quality checking routines and reports, and for the calculations of a measure to inform conceptual consistency according to INSPIRE standards. Besides, this structure allows exception handling but limited to internal datasets to which a field can be added defining if a specific feature is an exception or not to a particular rule.

   Nevertheless, the results showed that this format is not suitable for communicational purposes as there was found also evidence confirming that the standardized terminology is not universally understood, as was stated by Stock and Cialone (2011), but in this case related to the term “within” which was considered not enough clear to distinguish between features that are entirely or partly inside another feature. Deeper research is needed to confirm that this particular ambiguity is effectively present in the English-speaking population but, for the

purpose of this project, it suggests that natural expressions allow to include clarifying expressions when the standardized terminology seems confusing with the everyday use of the terms in the context of a particular pair of features. For this reason, the storage of spatial constraints as metadata based exclusively on the formalized structure seems not adequate when these metadata pretend communicational value among non-experts as the case of internal LBH data users. This finding limits the communicational usefulness of storing formalized constraints as metadata considered by Cockcroft (2004) and Von Oosterom (2006) as this would be only true for users with certain expertise in the formalized terminology, which is not the case in the LBH and probably in other similar contexts.

On the contrary, the results suggest that natural language and formalized expressions have opposed –or complementary– advantages and limitations that make suggest that a model to facilitate the automation of spatial DQ rules validation based on metadata should consider a combination of both languages. As a proposal, this can be done including the natural language expression of the rule along with the formalized version, the latter expressed as the answer to the question "which features pass the rule?", as shown in the example below. This would require to be tested in further research, ideally incorporating the perspective of other internal users such as the data custodians.

> "*Private roads should be wholely outside of public highways.*"
>
> So, which features pass the rule?
>
> Private road *crossing* = 0 AND *within* = 0 public highway

3. The results confirmed that the formalized version of the constraints is not automatically derivable from their natural language expression as the latter present ambiguities regarding cardinality, fields in referential external datasets and the meaning of the terms used to express topological relations so a manual translation process is needed. Therefore, this approach is limited by the fact that while it does not require scripting language, it does require certain knowledge about standardized terminology to refer to topological relationships which is required for the expression of constraints in a machine-readable format and cannot be assumed as universally understood, as was confirmed in the present research.

   The methodology used in the current research supposed this translation process done by the author and showed that the difference between the mathematical meaning of the standardized terms and their everyday use also affect it, so a method was suggested to define the topological relations implied by each expressed rule. However, it is not realistic to assign

this task entirely to the GIS team for all the constraints of all the LBH datasets. For this reason, it seems reasonable to incorporate the translation process to the workshops including a specific training about the mathematical meaning of the standardized predicates along with the method proposed to check which of them are implied by the expressed rule (*Section 5.2*), as a training to give the DOs the tools to express their rules in its formalized version and maintain them in case of changes.

Additionally, this method can be considered as part of the identification of rules itself as new rules can emerge from it because it makes one asks about all the possible relations together, helping to have a more consistent set of rules for a relevant pair of features. Even this is presumably easier to train than scripting languages, further research is required to determine a way to train non-experts in this topic, in an efficient and effective way. Depending on the required investment to train data managers in this topic and considering the high turnover existent among DO in the LBH, it would be possible to define the real applicability of the approach in this context or other similar organizations.

This new approach should also consider that DOs less familiarized with the use the team give to the data presents difficulties identifying rules required for it, so in these cases could be necessary to add other members of their departments.

4. Both types of reports proposed –considering statistics, general distribution map and graphical examples– were considered useful by the DO as a way to raise a problem and reinforce the communication of spatial constraints among internal users. They confirmed also that the identification of different types of issues (more than severity levels) allowed considering different solutions (including automatic solutions for some types of issues) and improved the utility of the reports. In general, the use of buffers showed more advantages to identify, in particular, coordinate precision problems (especially relevant when dealing with the relation "equals" between features from different sources) and the lack of snapping tools when defining features manually, because it is not affected by the size of the features as the percentage approach. However, apart from these issues, the cases analysed in detail are not enough to determine a complete model and suggest, on the contrary, that the specific workflow for these deeper analyses are strongly dependant on each specific case, thus limited in terms of scalability. Anyway, when a rule implies more than one topological relation, it seems convenient to check them separately and distinguish them in the report.

5. Finally, it should be considered that while it was not the focus of this project, non-binary topological relations were also relevant for the DOs and cannot be covered by the approach proposed.

Chapter 7: Conclusion

The research proved the relevance of data quality check routines and reports based on metadata as a suitable approach for spatial data quality management regarding logical consistency in the context of organizations where data management involves non-experts without the technical scripting knowledge required by standard approaches. Additionally, this approach appeared especially convenient to handle exceptions, which, contrary to the traditional definition of constraints in RDBMS, were transversally reclaimed.

But this approach is limited by the fact that while it does not require scripting language, it does require certain knowledge about standardized terminology to refer to spatial relation which is required for the expression of constraints in a machine-readable format and cannot be assumed as universally understood. The structure recommended by the literature to express these constraints proved to be adequate for the automation of these quality check routines and even for certain precision problems and exception handling, but not for communicational purposes. As it was confirmed, the standardized terms recommended by the international standards to refer to topological relations are not universally understood in accordance to it unambiguous mathematical meaning, so a combined approach was proposed to document binary topological spatial constraints as metadata. New research is required to test this approach, in particular its potential to communicate how the data is expected to be to internal data users and other potential users.

Finally, the reports successfully closed the circle going back to DOs with information that was considered useful to identify the problematic issues but as well as a tool to communicate data quality objectives itself to other internal users. The attempts to add more detail to these reports were successful for some specific kind of issues (coordinate precision problems and lack of snapping tools to define features) and were considered useful but other kinds of issues showed to be too dependent on each specific case so less scalable.

References:

Booch, G., Rumbaugh, J. & Jacobson, I.., 2005. *Unified Modelling Language User Guide*. [Online] 2nd ed. Reading, Massachusetts Addison Wesley. Available from: https://pdfs.semanticscholar.org/fc51/1dcebd3dae76133d5dbbda4250bebd0fb5e3.pdf [Accessed 02 08 2018]

AGI (Association for Geographic Information), 2012. *UK Gemini. Specification for discovery metadata for spatial data resources.* [Online] *v2.2.* London: AGI Standards Committee. Available from: https://www.agi.org.uk/about/resources/category/81-gemini?download=18:gemini-2-2 [Accessed 25 06 2018]

Atzeni, P., Ceri, S., Paraboschi, S. & Torlone, R., 1999. *Database Systems. Concepts, Languages, and Architectures.* [Online] London: McGraw-Hill. Available from: http://dbbook.dia.uniroma3.it/index.html [Accessed 20 08 2018]

Bogorny, V., Engel, P. & Alvares, L., 2005. *Towards the reduction of spatial joins for knowledge discovery in geographic databases using geo-ontologies and spatial integrity constraints.* [Online] Porto, Portugal, s.n. Availablre from: https://www.researchgate.net/publication/228625342_Towards_the_reduction_of_spatial_joins_for_knowledge_discovery_in_geographic_databases_using_geo-ontologies_and_spatial_integrity_constraints [Accessed 02 07 2018]

Borges, K., Davis Jr., C. & Laender, A., 2002. Integrity constraints is spatial databases. In: *Database integrity: challenges and solutions.* [Online] Portland: Idea Group Publishing, pp. 144-171. Available from: https://pdfs.semanticscholar.org/0ae9/ba772dd971d8510e7e4d5c64df6396cec1ef.pdf [Accessed 02 06 2018]

BSI (British Standardization Institute), 2006. *BS EN ISO 19125-1:2006 Geographic information - Simple feature acces - Part 1: Common architecture.* London: BSI Standards Limited.

BSI (British Standardization Institute), 2015. *BS EN ISO 19101-1:2014 Geographic Information. Reference model. Fundamentals.* London: BSI Standards Limited.

BSI (British Standardization Institute), 2018. *BS EN ISO 19157:2013+A1:2018 - Geographic information. Data Quality.* [Online] s.l.:BSI Standards Limited 2018.

Castells, M., 2010. *The rise of the network society.* [Online] 2nd ed. s.l.:Chichester : Wiley-Blackwell. Available from: https://deterritorialinvestigations.files.wordpress.com/2015/03/manuel_castells_the_rise_of_the_network_societybookfi-org.pdf [Accessed 02 08 2018]

Chrisman, N., 2006. Chapter 1. Development in the treatment of spatial data quality. In: R. Devillers & R. Jeansoulin, eds. *Fundamentals of spatial data quality.* [Online] s.l.:ISTE Ltd. Available from: https://onlinelibrary-wiley-com.libproxy.ucl.ac.uk/doi/pdf/10.1002/9780470612156 [Accessed 29 05 2018]

Clementini, E. & Di Felice, P., 1996. A model for representing topological relationships between complex geometric features in spatial databases. *Information Sciences,* [Online] 90(1-4), pp. 121-136. Available from: https://doi.org/10.1016/0020-0255(95)00289-8 [Accessed 02 06 2018]

Cockcroft, S., 1997. A taxonomy of spatial data integrity constraints. *GeoInformatica,* [Online] 1(4), pp. 327-343. Available from: https://link.springer.com/article/10.1023%2FA%3A1009754327059 [Accessed 02 06 2018]

Cockcroft, S., 2004. The Design and Implementation of a Repository for the Management of Spatial Data Integrity Constraints. *GeoInformatica,* [Online] 8(1), pp. 49-69. Available from: https://doi.org/10.1023/B:GEIN.0000007724.37467.ae [Accessed 02 06 2018]

Devillers, R. & Jeansoulin, R., 2006. Chapter 2. Spatial data quality: Concepts. In: *Fundamentals of Spatial Data Quality.* [Online] London: Iste Ltd, pp. 31-42. Available from: https://onlinelibrary-wiley-com.libproxy.ucl.ac.uk/doi/pdf/10.1002/9780470612156 [Accessed 29 05 2018]

Egenhofer, M. & Herring, J., 1994. Categorizing Binary Topological Relations between regions, lines and points in geographic databases. In: M. Egenhofer, D. Mark & J. Herring, eds. *The 9-Intersection: Formalism and its use for Natural language spatial predicates.* [Online] Santa Barbara, USA: NCGIA. Available from: https://escholarship.org/uc/item/5nj6647c [Accessed 02 06 2018]

ESRI, n.d. *ESRI GIS Dictionary.* [Online]
Available from: https://support.esri.com/en/other-resources/gis-dictionary/search/
[Accessed 15 08 2018].

FME, n.d. *Spatial relations defined.* [Online]
Available from:
http://docs.safe.com/fme/2018.0/html/FME_Desktop_Documentation/FME_Transformers/Transformers/spatialrelations.htm#DE9IM_Matrix
[Accessed 15 06 2018].

JRC (Joint Research Centre) of the European commission, 2013. *Data quality in INSPIRE: Balancing legal obligations with technical aspects.* [Online] s.l.:Publications Office of the European Union. Available from: http://publications.jrc.ec.europa.eu/repository/bitstream/JRC83209/lb-na-26-097-en-n.pdf [Accessed 06 06 2018]

Longley, P., Goodchild, M., Maguire, D. & Rhind, D., 2015. *Geographic Information Science & Systems.* [Online] 4th ed. s.l.:John Wiley & Sons. Available from: https://app.knovel.com/hotlink/toc/id:kpGISSE001/geographic-information/geographic-information [Accessed 02 08 2018]

Mark, D. & Egenhofer, M., 1994. Modeling Spatial Relations Between Lines and Regions: Combining Formal Mathematical Models and Human Subjects Testing. In: M. Egenhofer, D. Mark & J. Herring, eds. *The 9-Intersection: Formalism and Its Use for Natural-Language Spatial Predicates.* [Online] s.l.:s.n. Available from: https://escholarship.org/uc/item/5nj6647c [Accessed 02 06 2018]

Mark, D. & Egenhofer, M., 1995. *Topology of prototipical spatial relations between lines and regions in english and spanish.* [Online] Charlotte, North Carolina, USA, s.n., pp. 245-254. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.26.3560&rep=rep1&type=pdf [Accessed 02 06 2018]

Open Geospatial Consortium, 2011. *OpenGIS® Implementation standard for geographic informatin - Simple feature access - Part 1: Common architecture.* [Online] *Version 1.2.1.* s.l.:s.n. Available from: http://www.opengeospatial.org/standards/sfa [Accessed 18 06 2018]

Oracle, n.d. *Data Warehousing Concepts.* [Online]
Available at: https://docs.oracle.com/cd/A97630_01/server.920/a96520/concept.htm
[Accessed 25 July 2018].

Ordnance Survey, 2017. *OS MasterMap Topography layer technical specifications v2.0.* [Online]
Available at: https://www.ordnancesurvey.co.uk/docs/technical-specifications/os-mastermap-topography-layer-technical-specification.pdf
[Accessed 12 July 2018].

Preece, J., Sharp, H. & Rogers, I., 2015. *Interaction Design: Beyond Human-Computer Interaction.* 4th edition ed. Chichester: Jhon Wiley & Sons Ltd.

Riedeman, C., 2004. *Towards usable topological operators at GIS user interfaces.* [Online] Crete, Greece, Crete University Press, pp. 669-674. Available from: https://agile-online.org/conference_paper/cds/agile_2004/papers/8-1-3_riedemann.pdf [Accessed 02 07 2018]

Safe Software, n.d. *FME | Data Integration Platform | Safe Software.* [Online]
Available at: https://www.safe.com/how-it-works/
[Accessed 01 08 2018].

Servigne, S., Ubeda, T., Puricelli, A. & Laurini, A., 2000. A methodology for spatial consistency improvement of geographic databases. *GeoInformatica,* [Online] 4(1), pp. 7-34. Available from: https://link-springer-com.libproxy.ucl.ac.uk/article/10.1023/A%3A1009824308542 [Accessed 18 06 2018]

Smith, M., Goodchild, M. & Longley, P., 2018. *Geospatial analysis : a comprehensive guide to principles, techniques and software tools.* [Online] 6th edition ed. London: Leicester : Matador. Available from: www.spatialanalysisonline.com [Accessed 01 08 2018]

Stock, K. & Cialone, C., 2011. Universality, Language-Variability and Individuality: Defining Linguistic Building Blocks for Spatial Relations. In: M. Egenhofer, N. Giudice, R. Moratz & M. Worboys, eds. *Spatial Information Theory.* [Online] *COSIT 2011. Lecture Notes in Computer Science.* Berlin, Heidelberg: Springer. Available from: https://doi.org/10.1007/978-3-642-23196-4_21 [Accessed 02 07 2018]

Ubeda, T. & Servigne, S., 1996. *Geometric and topological consistency of spatial data.* Leeds, U.K., s.n., pp. 830-842.

Vallieres, S., Brodeur, J. & Pilon, D., 2006. Chapter 9: Spatial integrity constraints: a tool for improving internal quality of Spatial data. In: R. Devillers & R. Jeansoulin, eds. *Fundamentals of Spatial data quality.* [Online] s.l.:ISTE Ltd, pp. 161-178. Available from: https://onlinelibrary-wiley-com.libproxy.ucl.ac.uk/doi/pdf/10.1002/9780470612156 [Accessed 29 05 2018]

Van Oosterom, P., 2006. Chapter 7 Constraints in Spatial Data Models, in a Dynamic Context. In: J. Drummond, ed. *Dynamic and Mobile GIS: Investigating Changes in Space and Time.* [Online] Boca Raton: CRC Press, pp. 104-137. Available from: https://www.researchgate.net/publication/237462411_Chapter_7_Constraints_in_Spatial_Data_Models_in_a_Dynamic_Context [Accessed 02 06 2018]

Xu, D., Van Oosterom, P. & Sisi, Z., 2017. A methodology for modellin of 3D Spatial constraints. In: A. Abdul-Rahman, ed. *Advances in 3D Geoinformation.* [Online] s.l.:Springer, pp. 95-117. Available from: https://link.springer.com/book/10.1007/978-3-319-25691-7#about [Accessed 03 06 2018]

Appendices

Appendix 1: Information sheet and Informed consent form

Participant Information Sheet

YOU WILL BE GIVEN A COPY OF THIS INFORMATION SHEET

Title of Study:
DESIGNING AN APPROACH FOR AUTOMATIC DATA QUALITY CHECKING BASED ON METADATA

University/Department:
UNIVERSITY COLLEGE LONDON - CIVIL, ENVIRONMENTAL AND GEOMATIC ENGINEERING

Name and Contact Details of the Researcher(s):
Sofía Covarrubias Email: ucesova@ucl.ac.uk Tel: +44 (0) 7713 817577

Name and Contact Details of the Principal Researcher:
Claire Ellul Email: c.ellul@ucl.ac.uk Tel: +44 (0) 20 7679 118

Before you decided to participate in this study, it is important for you to understand why the research is being carried out and what participation will involve. Please take time to read the following information carefully and discuss it with the researcher if you feel you require additional information.

If you do decide to take part, you will be given this information sheet to keep and be asked to sign a consent form. You can withdraw at any time without giving a reason. If you decide to withdraw, provided your contribution was not kept anonymous, you will be asked what you wish to happen to the data you have provided up to that point.

The London Borough of Hackney's Geographical Information Systems team has implemented an extensive spatial data infrastructure (SDI) to manage the spatial (digital map) data held within the Borough. This SDI is built using FME software and an Oracle Spatial database but also, equally importantly, with in-house data management and data quality procedures. These data quality procedures should be based on spatial constraints (rules that describe how the data should be structured and how different datasets should relate to each other). These rules will be stored as metadata ('data about the data') from where checking/data quality validation procedures can be then implemented in the different subsystems of the SDI.

Expert knowledge of the datasets and what they are used for is required to define these rules – and this knowledge is held by the Data Owners. A process is therefore needed to facilitate the capture of these rules in a systematic way, to ensure that the overall quality of the data in the SDI can be maintained. In this context, the London Borough of Hackney's GIS team had established a collaboration agreement with the Department of Civil, Environmental and Geomatic Engineering at University College of London to develop a project (Masters dissertation) with a focus on spatial data quality.

The aim of the project is to explore and test ways to improve the definition of metadata about spatial constraints, and the use of it to create potentially automatable data quality routines and reports, driven by data user's requirements in the London Borough of Hackney GIS data management system. Specifically, this project will look at how the constraints/rules can be defined for input into the SDI and how a report on the data quality, which shows whether the data meets these rules, should be structured to be easily understood by data owners. This project is a pilot study which, if successful, maybe rolled out to a wider group at a later stage.

Inclusion Criteria
As a data owner, with responsibility for the content of certain Hackney's datasets, your knowledge, experience and opinion is really important for us so we are asking you to take part in this workshop, which may be followed up with an individual interview.

The workshop will take about 90 minutes. It will be focused on the definition of the required spatial constraints and will be accompanied by a preparation questionnaire. The specific objectives of the workshop are as follows:

Specific objectives:

1. Identify the most suitable way to enable the required spatial constraints to be defined by data owners, including specialists and non-specialists (this will be possibly based on a visual approach)

2. Work with data owners involved in this pilot study to define the quality requirements of their data, based on both the asset management needs of their departments.

How the Information will be Used
Based on the workshop results, a series of quality checking routines will be defined and a data quality report will be proposed for one or two of the participant departments. We therefore plan to use the outputs of the workshop as follows:

1. Translate the pilot data owners' quality requirements into spatial integrity constraints, looking at how this can be automated where possible.

2. Propose quality checking routines taking the defined quality metadata as an input. These routines should be doable using FME but the automation process will be not part of these research.

3. If time allows, define the content that should be included in an automatable report for data owners to show the constraints that exist on their data, and other information relating to the quality of the data, with a focus on generating a user-friendly report. The contents of this report will then be reviewed with data owners via a separate meeting.

4. As this project is part of a dissertation, the information will also be form part of the dissertation report.

Benefits to Participation
While you will not benefit personally from participating in this project, should the research be successful there will be direct benefits to the overall quality of the datasets for which you are responsible. This will not only help the end users of the data accomplish their work more successfully but could lead to an overall improvement in the quality of spatial data held within the London Borough of Hackney.

Please note that data and information collected during the course of the project might be used for additional or subsequent research and publications.

All contributions will be captured anonymously and no personal information will be captured during this project. However, you should note that it may be possible to indirectly link you to the information provided as any examples relating to the datasets for which you are responsible could be linked to you.

Thank you for reading this information sheet and for considering to take part in this research study.

Data Protection Privacy Notice

Notice:
The data controller for this project will be University College London (UCL). The UCL Data Protection Office provides oversight of UCL activities involving the processing of personal data, and can be contacted at data-protection@ucl.ac.uk. UCL's Data Protection Officer is Lee Shailer and he can also be contacted at data-protection@ucl.ac.uk.

Your personal data will be processed for the purposes outlined in this notice. *The legal basis that would be used to process your personal data is that your personal data (email contact details) will only be used by the researcher to invite you to participate in follow on exercises and/or to identify*

the data you provide, should you wish to withdraw at a later stage. You can provide your consent for the use of your personal data in this project by completing the consent form that has been provided to you.

If you are concerned about how your personal data is being processed, please contact UCL in the first instance at data-protection@ucl.ac.uk. If you remain unsatisfied, you may wish to contact the Information Commissioner's Office (ICO). Contact details, and details of data subject rights, are available on the ICO website at: https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/individuals-rights/

CONSENT FORM FOR *COUNCIL EMPLOYEES* IN RESEARCH STUDIES

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

Title of Study: Designing an approach for automatic data quality checking based on metadata
Department: University College London - Civil, Environmental and Geomatic Engineering
Name and Contact Details of the Researcher(s):
Sofía Covarrubias, ucesova@ucl.ac.uk, Tel: +44 (0) 7713 817577
Name and Contact Details of the Principal Researcher:
Claire Ellul, c.ellul@ucl.ac.uk, Tel: +44 (0) 20 7679 118
Name and Contact Details of the UCL Data Protection Officer:
Lee Shailer, data-protection@ucl.ac.uk

Thank you for considering taking part in this research. The person organising the research must explain the project to you before you agree to take part. If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.

I confirm that I understand that by ticking/initialling each box below I am consenting to this element of the study. I understand that it will be assumed that unticked/initialled boxes means that I DO NOT consent to that part of the study. I understand that by not giving consent for any one element that I may be deemed ineligible for the study.

		Tick Box
1.	*I confirm that I have read and understood the Information Sheet for the above study. I have had an opportunity to consider the information and what will be expected of me. I have also had the opportunity to ask questions which have been answered to my satisfaction and would like to take part in (please tick one or more of the following) - a preparation questionnaire for the joint interview/workshop and the workshop itself - an individual interview	☐ ☐ ☐
2.	*I understand that I will be able to withdraw my data for up to the 15th August 2018 (after which time the data will be processed by the project and will not be able to be withdrawn).	☐
3.	*I consent to the processing of my personal information *(name, role, contact details)* that will be used only to contact me and will not be linked with the data gathered. I understand that such information will be handled in accordance with all applicable data protection legislation.	☐ ☐
4.	**Use of the information for this project only** The information gathered in this project will be anonymized and all efforts will be made to ensure you cannot be identified. This entails that: • Your real name, position and contact details will not be in any way digitally linked to the information you provide. However, the nature of the data gathered make it impossible to completely remove any reference to its source, specifically because the departments participating in the projects may be derived from the spatial objects	

	considered in the analysis. As there is only one data owner per department, the information gathered from you may be connected to yourself. *I understand that even the data will be anonymized and no personal information will be captured as part of the project (beyond that required to contact you, which will be held completely separately from the project data), the information gathered from me may be connected with me indirectly.	☐
5.	*I understand that the information provided may be subject to review by responsible individuals from the University for monitoring and audit purposes.	☐
6.	*I understand that my participation is voluntary and that I am free to withdraw at any time up to the 15th August 2018, without giving a reason (after this point, the information will be processed for the dissertation). I understand that if I decide to withdraw, any data I have provided up to that point will be deleted unless I agree otherwise.	☐
7.	I understand the indirect benefits of participating.	☐
8.	I understand that the data will not be made available to any commercial organisations but is solely the responsibility of the researcher(s) undertaking this study.	☐
9.	I understand that I will not benefit financially from this study or from any possible outcome it may result in it the future.	☐
10.	I agree that my anonymised research data may be used by others for future research.	☐
11.	I understand that the information I have submitted will be published as a report and I wish to receive a copy of it.	☐ Yes ☐ No
12.	I hereby confirm that I understand the inclusion criteria as detailed in the Information Sheet and explained to me by the researcher.	☐
13.	I am aware of who I should contact if I wish to lodge a complaint.	☐
14.	I voluntarily agree to take part in this study.	☐
15.	Use of information for this project and beyond: I would be happy for the data I provide to be archived at Hackney's repositories and at University College London, where it may be used for subsequent research. I understand that other authenticated researchers or Hackney Staff may have access to my anonymised data.	☐ ☐

If you would like your contact details to be retained so that you can be contacted in the future by UCL researchers who would like to invite you to participate in follow up studies to this project, or in future studies of a similar nature, please tick the appropriate box below.

	Yes, I would be happy to be contacted in this way	
	No, I would not like to be contacted	

Name of participant	Date	Signature
Researcher	Date	Signature

Appendix 2: GIS training and experience online survey

Training and experience in GIS

Your email address (**sofia.covarrubias@hackney.gov.uk**) will be recorded when you submit this form. Not **sofia.covarrubias**? Sign out
*Required

1. **Do you have any training in geographic information science? please select between the following:** *
   Mark only one oval.
   - ◯ No training in gis, just practical experience *Skip to question 8.*
   - ◯ Modules in an university degree (not gis degree) *Skip to question 2.*
   - ◯ A gis degree (completed or not completed) *Skip to question 3.*
   - ◯ Formal training programme in gis (not university) *Skip to question 5.*
   - ◯ Informal studies *Skip to question 6.*
   - ◯ Other *Skip to question 7.*

2. **How many modules in GIS did you have?**
   Mark only one oval.
   - ◯ 1 to 2
   - ◯ 3 to 4
   - ◯ More than 4

Skip to question 8.

3. **Did you have completed this degree**
   Mark only one oval.
   - ◯ Yes *Skip to question 8.*
   - ◯ No

4. **How many modules have you done so far? (provide a number)** *
   _____

Skip to question 8.

5. **Can you specify what kind of training was this and how long it was?**
   _____
   _____
   _____
   _____
   _____

Skip to question 8.

6. **Can you specify what kind of studies were these?**

Skip to question 8.

7. **Can you specify?**

Skip to question 8.

Work experience

8. **How many years have you been working on things related with GIS? Please answer with a number** *

Summary of the results:

One of the pilot data owners do not have any training in GIS, just practical experience of 4 years working in related topics, three of them had 1-2 modules about GIS in a university degree (not GIS degree), plus more than 10 years of experience in the field, and one is currently studying a GIS degree in its first semester with one year of experience in the field.

Appendix 3: Preparation questionnaire

Note: Page three was removed as was an empty page to be completed by the participant.

PREPARATION QUESTIONNAIRE
"Defining what does the team needs from the data"

In partnership with UCL GIS MSc programme, Hackney GIS team is exploring ideas and tools related to spatial data quality. The research questions are:

- What are our spatial data quality constraints?
- Could data owners express these constraints in the metadata of their datasets?
- Using the expressed constraints, could we run automatic quality checks and derive useful data quality reports for data owners?

This questionnaire is an information gathering exercise that will pave the way for our interviews/workshops planned for next week, where we will start defining together what is "good quality data" in the context of our job in Hackney.

One way to understand data **quality** is **how it is able to fulfil specific requirements**. This is sometimes called as *fit-to-purpose* and the first step to evaluate it is to explicit what the user needs to do with the data. This is what we want to ask you now, as a preparation for the interviews next week. With these requirements specified, it will be then possible to define the way the data should be stored to be useful for that needs, and to define measurements to assess the four broad quality components: *completeness*, *consistency*, *accuracy (positional, temporal and thematic),* and *usability*.

The questionnaire is divided in two general questions, the first one, related to Hackney's **operational needs** while the 2nd with **general quality considerations** in order to make the data clean, easy to understand (either by Hackney's users and external potential users), republishable, or used to feed other systems without crashing .

We add some examples for each question, as hypothetical situations about what you and your team would like from your datasets. Please consider that these examples can make no sense to your particular datasets. They are just to illustrate the kind of things your team can think about. Don't consider yet how easy these things are, or how they can be translated into quality requirements. Just consider what kind of information you want to obtain from the dataset, or what you expect from it to be considered good-quality datasets from your perspective.

Thank you very much!

1

Question 1: What are your team requirements related to operational needs? (if it helps, formulate them as questions you would want to answer using your datasets)
Example 1: ***A new regulation requires that all playgrounds in public spaces and educational institutions should have an impact-protection surface in the floor, covering the whole area.*** To estimates the cost of the implementation of this regulation, this situation entails asking: • Which are the public spaces and educational institutions managed by the council that have playgrounds? • What is the area of these playgrounds? Example 2: ***There is a plague affecting a certain type of trees in London that makes them extremely week provoking the fall of branches. The borough's estates maintenance team wants to fumigate them starting by the trees near to parking spaces and playgrounds, to prevent major damage.*** This situation entails making the following question to the data: • Which estates have this specific type of tree? • Which estates have this specific type of tree planted next to parking spaces and playgrounds? Example 3: ***To maintain Hackney's streets, street cleaning management areas were defined and they should cover all the territory.*** *A question to be done would be:* • *Is there any gap in the territory not covered by a street cleaning management areas?* • *Is there any part of the territory covered by more than one of these areas?*
Your answer:

2

Question 2: What are your team requirements related with general quality considerations?
Example 1: • ***The planning software does not allow self-intersected polygons (i.e. a polygon that make a loop over itself, making an edge of the polygon crosses another edge). To avoid potential problems this kind of polygons should not be allowed in layers used as planning constraints.*** Example 2: • ***A specific layer is shared with the GLA and republished on the London Data Store. The map uses Ordnance Survey Master Map as a background. The map will look clearer, and the data will be easier to reuse, if the layer's geometry follows exactly lines and polygons from the map background (features referring to the same object in the real world should be coincident, independent of the dataset).***
Your answer:

4

Appendix 4: Workshop outline

WORKSHOP PLAN
How a quality dataset looks like for each of us
Time: 95 minutes

OUTLINE:

• Presentation	15 minutes	
• Main activity:		
○ Instructions	5 minutes	
○ Individual work	15 minutes	
○ Presentation and discussion:		
■ Participant 1	15 minutes	
■ Participant 2	15 minutes	
■ Participant 3	15 minutes	
• Final thoughts:	15 minutes	

WORKSHOP OBJECTIVES :

1. Identify what kind of rules need to be encoded in the metadata (necessary for the project)
2. See how people express these rules, so we can:
    a. give them an intuitive interface (out of scope) to fill the metadata.
    b. have an idea of how to re-explain these rules in the quality reports, which are aimed at different audiences.

PLAN:

Presentation:
Time: 15 minutes
(See appendix 4 for supporting presentation)

1. The responsible of the project broadly presents the project and the objectives of the workshop.
2. Explain the definition of quality that we are following and that the preparation questionnaire was to identify the "purpose" of the data.
3. Explain the main quality elements.
4. Explain that the focus is in spatial requirements. If there are others requirements comment them anyway but briefly
5. Give space for each participant to present him/herself and answer the questions in slide 3:
    a. What do you think about about expressing these rules explicitly?
    b. And about generating reports based on these rules?

Main activity: Identifying the rules that are behind our idea of a good enough dataset
Time: 65 minutes
(5 for instructions and 15 minutes for individual work, 15 for each participant to explain the rules identified and receive comments)

Materials:

- Projector (projecting Earthlight)
- Notebook for each participant to allow them access to Earthlight
- Mouse for each
- A set of 20 printed forms A for each participant (1 color per participant) (see appendix 5)
- A set of markers of 5 different colours and a pencil for each participant
- Eraser
- Tip-ex
- Blu tack

Before start: explain them that the activity will be recorded

Instructions:
Imagine that you need to explain to a new data custodian (with limited experience GIS) how the information should be represented in the datasets. Your purpose is to make sure that the relevant objects are correctly represented in the map reducing the most possible the editing errors, especially in the features that are more important for the department's needs. We ask you to define this as a list of "rules", writing and drawing them in the form A (one rule per form). Write in the back any explanation or comment you think is important and draw it in the way you think is the most clear to understand the rule. At the end we will share what all of us did.

To do this, we will give you 15 minutes to check your datasets in Earthlight trying to identify the rules that are behind your idea of a "good enough" dataset (for your purposes).

Some helping questions to make yourself are:

- *Which are the most relevant features? This can be derived from the preparation questionnaire*
- *Is there anything that looks wrong when checking these relevant features in Earthlight?*
- *What kind of thing could goes wrong in the data acquisition or editing process? Can a rule help to check this?*
- *Is (or could be) there any data duplicated or clearly omitted? Can a rule help to check this?*
- *Is (or could be) there any data on a too wrong position? Can we define forbidden locations for the features? Or certain distance to something else?*
- *Is there any implicit rule about the size or shape of these feature?*

After this, in turns, each of you will explain to the rest of us the rules that you had identified. Earthlight will be projected so you can support your explanation showing your datasets there.

(15 minutes for individual work with the "helping questions" slide projected)

Now lets share what you find in turns. The idea is explain and discuss them with the rest, and paste the forms with your rules on the spatial constraints poster (see supporting presentation, slide 6), next to the

corresponding spatial constraint type. If while listening to the others you think in some rules that you didn't write in the forms, you can just add them.

Steps:

1. 15 minutes of individual work
2. Participant 1 explains his/her rules pasting them in the spatial constraints "poster", next to the corresponding spatial constraint type.
3. Ask the rest of the participants if they have any question/comment/suggestion
4. Continue with the next participant

Final thoughts:

Time: 15 minutes

To close the workshop, explains what are the following steps in the project and ask the participants if they have any comment, conclusion or recommendation to continue with the process.
Thank all of them for their time.

Appendix 5: Workshop's supporting presentation

Designing an approach for automatic

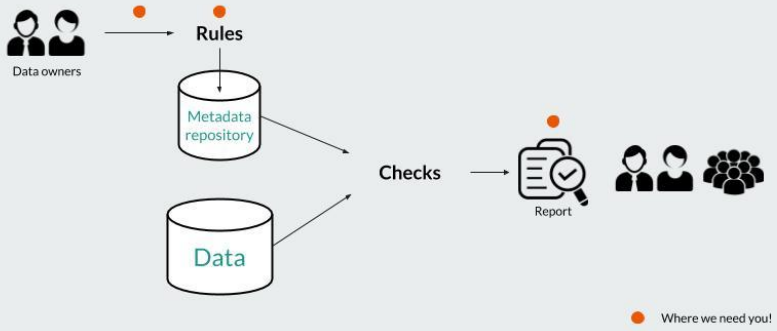
spatial data quality checking

based on metadata

1

Hackney UCL

Project outline:



2

Hackney UCL

Project outline:

- *What do you think about about expressing these rules explicitly?*
- *And about generating reports based on these rules?*

3

Hackney UCL

What is quality?

"Degree to which a set of inherent characteristics fulfils requirements"

INSPIRE

Hackney UCL

How to assess quality: Data quality elements

- Completeness:
- Logical consistency
- Positional accuracy
- Thematic accuracy
- Usability

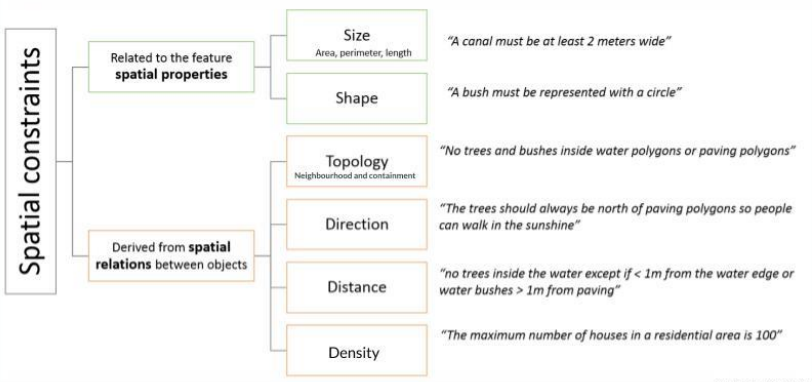
To evaluate the difference between:

a perfect dataset for the purposes of the users

INSPIRE, based on ISO 19157:2013 (Geographic Information - Data Quality)

Hackney UCL

Focus on:



(Based on Van Oosterom, 2006)

Hackney UCL

Guidance questions:

- *Which are the **most relevant features**? This can be derived from the preparation questionnaire*
- *What spatial queries can **answer our requirements**?*
- *Is there anything that **looks wrong** when checking these relevant features in Earthlight?*
- *or , **What** kind of things **could go wrong** in the data acquisition or editing process?*
  - → *Can a rule help to check this?*
- *Is (or could be) there any data **duplicated** or clearly **omitted**?*
  - → *Can a rule help to check this?*
- *Is (or could be) there any data on a too **wrong position**?*
  - → *Can we define forbidden locations for the features?*
  - → *Or certain distance to something else?*
- *Is there any implicit rule about the **size** or **shape** of these features?*

7

Hackney UCL

For example:

Purpose

Hypothetical requirements:

- Which are the LBH estates that don't have recycling bins?

Rules

- The estates recycling bins should be inside the estates polygons.

8

Hackney UCL

For example:

Purpose

Hypothetical requirements:

- Which are the LBH estates that don't have recycling bins?

Rules

- The estates recycling bins should be inside the estates polygons.
- The estates polygons should correspond to the whole parcel (not only the building), following the OS mastermap polygon

9

For example:

- Should be each type of light over a certain type of area?

FIG 1: Simmon signs illuminated LED sign posts in different type of areas. One over a road and the other over the roadside

FIG 2: Drop down lighting column (green dot) within a building

10

For example:

- Can a road be classified as "One way road" (dotted black line) and as "road" (normal black line) at the same time?

11

For example:

- If some asset should be always in pairs, for example combined illuminated zebra beacon posts, the lack of a pair may suggest an omission.

- There should be always one zebra beacon post in a distance no bigger than 6 meters from another zebra beacon post.

12

Hackney UCL

Workshop plan:

- Individual work → 15 minutes
- Sharing what you found → in turns, 45 minutes
- Final thoughts → 15 minutes

13

Hackney UCL

Guidance questions:

- *Which are the **most relevant features**? This can be derived from the preparation questionnaire*
- *What spatial queries can **answer our requirements**?*
- *Is there anything that **looks wrong** when checking these relevant features in Earthlight?*
- *or , **What** kind of things **could go wrong** in the data acquisition or editing process?*
  - → *Can a rule help to check this?*
- *Is (or could be) there any data **duplicated** or clearly **omitted**?*
  - → *Can a rule help to check this?*
- *Is (or could be) there any data on a too **wrong position**?*
  - → *Can we define forbidden locations for the features?*
  - → *Or certain distance to something else?*
- *Is there any implicit rule about the **size** or **shape** of these features?*

14

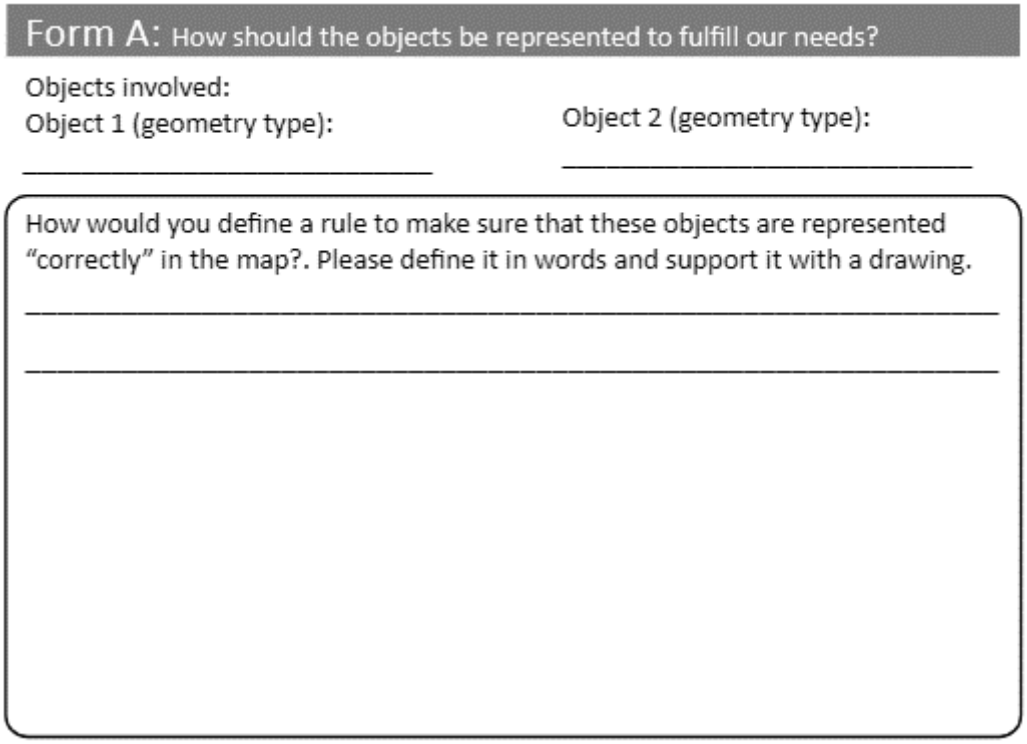
Hackney UCL

Final thoughts:

- *How do you think the definition of constraints can be eased for the data owners?*
- *How do you think the data constraints should be showed to the data custodians so they can follow them up?*
- *Any other comments?*

15

Appendix 6: Workshop's forms A to be filled by participants with spatial constraints



Form A: How should the objects be represented to fulfill our needs?

Objects involved:

Object 1 (geometry type): _____

Object 2 (geometry type): _____

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

_____

_____

Appendix 7: Outline of the report's feedback interview with the data owners

REPORT'S FEEDBACK INTERVIEW OUTLINE

Questions about basic reports:

1. In general, what do you think about a report like this:
    - Checking questions (make them only in case they are not answered in the first natural answer):
        - How useful is for you as a tool to improve the quality of your data?
        - What you would do with this data,
        - How you think you can correct these issues:
2. In the 2nd page, the examples are shown using different styles. Which one do you prefer and why?
    - Consider:
        - Presence of base map
        - Chart type
        - Presence of cartographic text in the base map
        - Reference layers color:
            - Grey (with transparency)
            - the color currently in use in Earthlight
            - Colour and size of the features being checked:
        - Highlighting only the part of the private road that is inside a public highway
        - Labels:
        - (check if there is anything specific for the layer being checked)
3. What you would like to change to the report to make it more useful or easy to understand?
4. Apart from the report, how you would like to receive the data of the features with issues?
5. Is the data inside the tables the one that you would want to see there?
6. What do you think about the alternatives suggested to solve the issue?

Questions about detailed reports:

(Give them a time to check the report alone, without explanation)

1

1. Do you think that a more detailed report like this improves the usefulness of the report as a tool to improve the quality of your data?
    - Checking question:
        - Does this change the way you can use the data and correct these issues?

2. What do you think about the idea of defining a “severity level”? And about the name we choose itself.

3. Do you agree with the criteria used to define the severity level?

4. What you would like to change to the report to make it more useful or easy to understand?
    - Checking question:
        - What do you think about the chart
        - Do you think the graphic way to express the criteria used to define the severity level is clear and easy to understand?

5. Apart from the report, how you would like to receive the data of the features with issues?
    - Checking question:
        - 1 tab file for all the features with issues with a severity level attribute or 1 tab file per each severity level?

7. What do you think about the alternatives suggested to solve the issue?

2

Appendix 8: Preparation questionnaire answers

a) Data owner nº1

Question 1: What are your team requirements related to operational needs?

- We require area measurements for properties owned by the Council, however the returned results source from the land registry layer does not accurately capture the correct measurements because the registered boundaries does not reflect the building demise
- Identifying building heights
- Land registry ownership boundaries capturing a building ownership regular extends onto the road and pavements. Clearly identifying Council owned properties boundaries would be beneficial.
- Clearly establishing hardstanding lands that are owned by the Council that does not form part of any building ownership or council management

Question 2: What are your team requirements related with general quality considerations?

- No answer

b) Data owner nº2

Not answered

c) Data owner nº3

Question 1: What are your team requirements related to operational needs?

- Which estates have recycling bins?
- Which are Hackney owned and privately-owned estates?
- What type of recycling bins on the estate? E.g. Food, Garden, Mixed
- How many bins on each estate?
- Number of properties vs number of bins
- Types of Anti-Social Behaviour (ASB) found on estates
- Create hot spots of ASB per estate then moving to a higher-level ward
- Type of issues experienced in Ward Improvement Programme
- Rank issues through hotspot or heat map
- Type of services received by residential properties based on property type (BLPU classification)
- Which properties are not shown as receiving a service?
- Which estates do not have a recycling bin – analyse by bin type (food, garden waste)?
- Which estate polygons do not contain a BLPU point?
- Which estate polygons do not contain a Food recycling point?
- Which schools are within a high pollution count zone?

Question 2: What are your team requirements related with general quality considerations?

- Layer should not contain multiple geometry types
- Consistency in data fields within the database – e.g. Day field should be abbreviated or full Tue/ Tues or Tuesday
- Identifying slivers less than a specific area
- Cross reference spatial data captured to OS MasterMap's base map
- Check for duplicate records or geometry

d) Data owner nº4

Not answered

e) Data owner nº5

Question 1: What are your team requirements related to operational needs?

- We need to know which buildings in the borough are listed, locally listed or in conservation Areas.
- We need to know the date, grade and extent of designation.
- We need to be able to see the information separately and together.
- Sometimes we need to see this information with other information e.g. which listed, locally listed or Conservation Area buildings belong to the Council or are in the LLDC area.

Question 2: What are your team requirements related with general quality considerations?

- There is an issue about how the GIS system deals with curtilage listing and the listing of unusual elements which are not buildings e.g. railings, war memorials. Where the listing is of a long thin object, this displays oddly. It is awkward if the object is not on the Ordnance Survey layer. The extent of curtilage listing is not currently displayed properly (this is a data input rather than a system or display issue).

Appendix 9: non-binary and non-topological rules that were expressed by the data owners

NON-BINARY TOPOLOGICAL RELATIONS THAT CANNOT BE BINARISED	
	The footway and carriageway should not overlap or have gaps between
	Red route/ LBH public highways should not overlap and not have gaps
OTHER NON-TOPOLOGICAL RELATIONS	
	Recycling estates should snap to OSMM lines
	All aerials must be presented as circles
	Crossing polygon and zebra beacon point should be within X meters of each other.
	Point data (general practitioners/opticians/pharmacies/hospitals/dentists/clinics) [should have] reasonable accurate location for visual display (needs flexibility)
	Other locations relevant to health determinants (point data) [should have] reasonable accurate [location] for visual display and calculation of densities, rates per head, travel times.
	Boundaries (Local Super Output Areas (LSOAS)/wards/grouped LSOAS/neighbourhoods) [should have] reasonable accurate [location] for visual display and calculation of densities, rates per head, travel times.

Appendix 10: Workshops forms filled by the data owners with the spatial rules for their data

(9) L

Form A: How should the objects be represented to fulfill our needs?

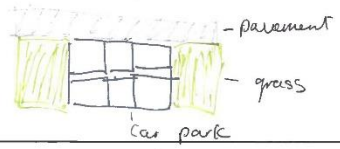
Objects involved:

Object 1 (geometry type): Car parks (Areas)

Object 2 (geometry type): Topo areas from layer map (Areas)

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

Car Parks areas should not include any topographic area covered with grass.



(12)

Form A: How should the objects be represented to fulfill our needs?

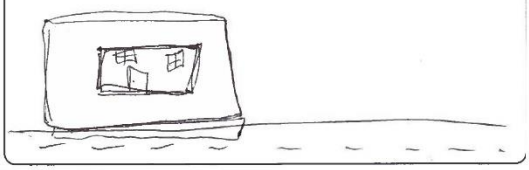
Objects involved:

Object 1 (geometry type): polygon

Object 2 (geometry type):

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

polygon identifying a building alone and the ownership boundary with no extension onto the Road



(11)

Form A: How should the objects be represented to fulfill our needs?

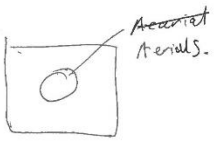
Objects involved:

Object 1 (geometry type): points

Object 2 (geometry type): ~~polygons~~ polygons

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

All Aerials must be presented as circles



(10) positional Accuracy L

Form A: How should the objects be represented to fulfill our needs?

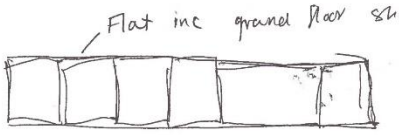
Objects involved:

Object 1 (geometry type): flats (Area)

Object 2 (geometry type): buildings (Area)

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

polygons identifying the demise of a shop.



(2) J

Form A: How should the objects be represented to fulfill our needs?

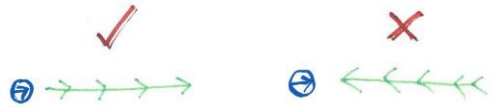
Objects involved:

Object 1 (geometry type): One way streets

Object 2 (geometry type):

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

depiction direction should be the same as the direction of the one way restriction



(1) P

Form A: How should the objects be represented to fulfill our needs?

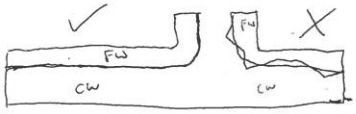
Objects involved:

Object 1 (geometry type): Footway POLY

Object 2 (geometry type): Carriageway POLY

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

the FW + CW should not overlap or have gaps between



(4) J

Form A: How should the objects be represented to fulfill our needs?

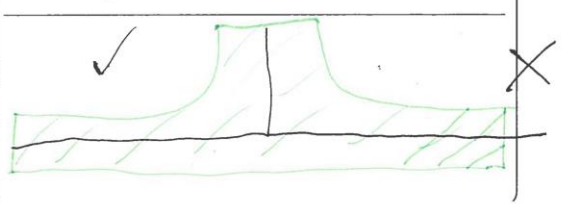
Objects involved:

Object 1 (geometry type):

Object 2 (geometry type):

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

LBH managed roads wholly with LBH public highway



(3) J

Form A: How should the objects be represented to fulfill our needs?

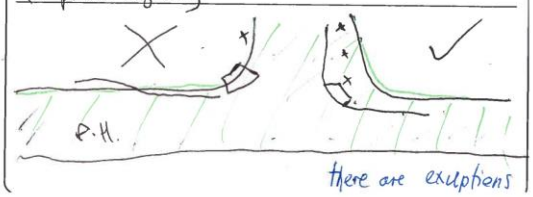
Objects involved:

Object 1 (geometry type): ASSETS (Point Line Poly)

Object 2 (geometry type): Public Highway

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

Highway assets should only appear within the polygon for public highway

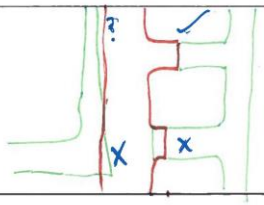


(6) J

Form A: How should the objects be represented to fulfill our needs?

Objects involved:
Object 1 (geometry type): LBH public Hwy
Object 2 (geometry type): TLRN

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.
Red route / LBH pub Hwy should not overlap

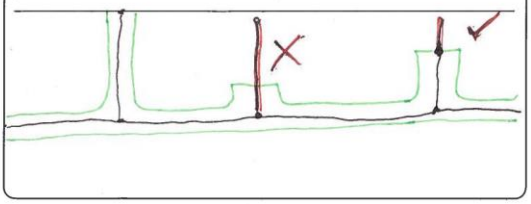


(5) J

Form A: How should the objects be represented to fulfill our needs?

Objects involved:
Object 1 (geometry type): private roads
Object 2 (geometry type): Public Highway

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.
private roads should be wholly outside of Pub Hwy

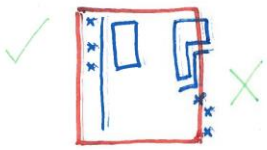


(8) J

Form A: How should the objects be represented to fulfill our needs?

Objects involved:
Object 1 (geometry type): Housing Assets
Object 2 (geometry type): Housing Estates

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.
All housing estate Assets should be within Housing estates polygon



(7) J

Form A: How should the objects be represented to fulfill our needs?

Objects involved:
Object 1 (geometry type): Zebra beacon
Object 2 (geometry type): Zebra crossing

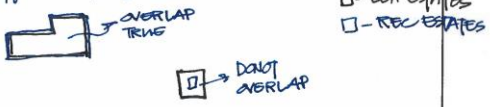
How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.
crossing polygon + Zebra beacon point should be within x metres of each other. NOT WITHIN

(3) B

Form A: How should the objects be represented to fulfill our needs?

Objects involved:
Object 1 (geometry type): POLYGON - LBH ESTATE
Object 2 (geometry type): POLYGON - REC ESTATES

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.
- Identify where polygons do not overlap
- Identify where polygons overlap



(1) B

Form A: How should the objects be represented to fulfill our needs?

Objects involved:
Object 1 (geometry type): POINT - REC BINS
Object 2 (geometry type): POLYGON - ESTATES

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.
Identify all REC BINS not within ESTATES or not within 5m of ESTATE

(4) S

Form A: How should the objects be represented to fulfill our needs?

Objects involved:
Object 1 (geometry type): OTHER LOCATIONS RELEVANT TO RESULTS DETERMINANTS
Object 2 (geometry type): POINTS

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.
eg: FAST FOOD PREMISES; LEISURE CENTRES; PARKS; PUBS; BETTING SHOPS; OFF-LICENCES; LIBRARIES; CULTURE; POLLUTION
REASONABLE ACCURATE FOR VISUAL DISPLAY & CALCULATING DENISTY / RATES PER HEAD. / TRIVIAL THINGS.

(3) S

Form A: How should the objects be represented to fulfill our needs?

Objects involved:
Object 1 (geometry type): POLYGON - REC ESTATE
Object 2 (geometry type): POINT - LLPG

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.
Identify REC ESTATES without RP classification not RD06 (Flats)

Form A: How should the objects be represented to fulfill our needs?

Objects involved:

Object 1 (geometry type): ~~BOUNDARIES~~ - LSOAs / WARDS

Object 2 (geometry type): GROUPED LSOAs - "NEIGHBOURHOODS"

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

REASONABLY ACCURATE FOR VISUAL DISPLAY + CALCULATIONS

SINGLE, JOINED, LINES

DENSITY / ROUTES PER WARD / TRAVEL TIMES

---

(5)

Form A: How should the objects be represented to fulfill our needs?

Objects involved:

Object 1 (geometry type): GP PRACTICES / OPTICIANS / PHARMACIES / HOSPITALS / DENTISTS / CLINICS

Object 2 (geometry type): POINTS

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

IS REASONABLY ACCURATE LOCATIONS (~~FOR~~ VISUAL DISPLAY)

SOME PRACTICES ARE "CO-LOCATED" OR ON 2 SITES.

NEEDS FLEXIBILITY TO SHOW THIS.

WITHIN BOROUGH BOUNDARIES.

---

(8)

Form A: How should the objects be represented to fulfill our needs?

Objects involved:

Object 1 (geometry type): Conservation Area

Object 2 (geometry type): OS Map

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

A Conservation Area should not cut a building in half.

---

(7)

Form A: How should the objects be represented to fulfill our needs?

Objects involved:

Object 1 (geometry type): Conservation Area

Object 2 (geometry type): OS Map

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

A building should not be in more than one Conservation Area.

---

(14)

Form A: How should the objects be represented to fulfill our needs?

Objects involved:

Object 1 (geometry type): Nationally or locally listed building

Object 2 (geometry type): OS Map.

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

A nationally or locally listed building should include data on
- asset type (eg. building or street furniture)
- ownership (eg. Council or private).

---

(16)

Form A: How should the objects be represented to fulfill our needs?

Objects involved:

Object 1 (geometry type): Nationally or locally listed

Object 2 (geometry type): OS Map

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

The shape of a building should be the shape on OS map
- with exceptions.

---

(12)

Form A: How should the objects be represented to fulfill our needs?

Objects involved:

Object 1 (geometry type): Nationally or locally listed

Object 2 (geometry type): OS Map.

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

The shape of a listed building should be as it appears on the OS map.

---

(6)

Form A: How should the objects be represented to fulfill our needs?

Objects involved:

Object 1 (geometry type): Nationally listed buildings

Object 2 (geometry type): Historic England

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

Hackney's data on nationally listed buildings does not exactly match Historic England data (and it should).

11

Form A: How should the objects be represented to fulfill our needs?

Objects involved:

Object 1 (geometry type): NATIONALLY LISTED

Object 2 (geometry type): LOCALLY LISTED

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

One building should not be both nationally and locally listed.

12

Form A: How should the objects be represented to fulfill our needs?

Objects involved:

Object 1 (geometry type): Conservation Area

Object 2 (geometry type): OS Map

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

A Conservation Area must be in Hackney

13

Form A: How should the objects be represented to fulfill our needs?

Objects involved:

Object 1 (geometry type): Nationally or locally listed

Object 2 (geometry type): OS Map

How would you define a rule to make sure that these objects are represented "correctly" in the map?. Please define it in words and support it with a drawing.

The shape of a nationally or locally listed building should show clearly the exteent of curtilage listing (walls, railings) etc.

Appendix 11: Transcription of the first workshop notes

Participants:

- DO1: Data owner number 1
- DO2: Data owner number 2
- GIS1: GIS team member number 1
- GIS2: GIS team member number 2
- Int: Interviewer

Note:

- The squared brackets correspond to comments of the author to try to explain things that were expressed in a non-verbal communication.

Transcription:

DO1: The issues that we experience are more associated with the accuracy because the buildings that we manage we want to visualise what the building extent is and the boundary separately. I'm not sure whether it is a constraint. For example, the town hall. Some people would like to know the ownership and it is split into two because of the land registry. We would want to simplify that and show the town hall its own and then the boundary separately. Would that be a constraint? Maybe is something that we need to work on before establishing the constraints?

GIS2: Maybe it would be a constraint for a future layer.

Int: Would this future layer follow something?

DO1: Yes, it would do, because another thing is what we would want to know is the size of the assets and trying to save the trouble of that finding it and drawing the polygon we just want to pull it from the polygon, but what is happening is that the polygon [in the land registry hackney owned layer] is not accurately capturing the building alone, is capturing also a road. It should be following exactly the basemap object but not the street, only the building.

In another example in Stoke Newington, it is a massive site which is connected to a library and the land registry includes the library's part of its demise, which is separate from what we manage. And we would need that "gaps" to be incorporated

GIS1: so, you need the parcel with the garden. That would be a different layer, I think.

DO1: Yes, because if there is a property with a garden, how would you know if that garden is part of the ownership of that building and that we need to manage it?

Int: and how you would like that polygons to be. Is there any rule you can define for them?

DO1: they should not go into the road. I put: "A polygon identifying a building alone, and then the ownership boundary without extension into the road" because sometimes the boundary goes to that pavement or roads and then what happens if we want to see what we own and what we are responsible for maintaining.

DO2: Mine is about carriageways and footways. They should not overlap, and they shouldn't have gaps between the two. So, it is topology.

DO1: We also manage aerials in the estates and they must be represented as a circle.

DO2: But it should be a point or a polygon?

DO1: a point

DO2: This is about multiple layer: Highway assets should only appear within the public highway polygon, most of the time. There are some exceptions but most of the time it is like that.

GIS1: But, how do you define that exceptions? is there any special kind of asset that doesn't need to follow the rule? Is it possible to say something like that?

DO2: No. I just know where the exceptions are. I can't think of a rule for the exceptions.

GIS1 So it is valuable to check and the most of them are inside? The polygon layer covers the pavement, as well?

DO2: Yes, some of the time. Sometimes the footway is really weird but that's how it is.

GIS2: I have one: There are some lines of very specific categories, entry gates, fences, that needs to be matching the boundaries of the polygons, the park in this case. I think is topology but shape as well because fences should be lines.

DO2: I have one about distance I think: zebra crossing polygons on the road and the beacons, they should be within certain distance, and not within the crossing and not too far.

GIS1: Do you think this rule could help someone to do something? Because we know how you would like the data to be but how important it is?

DO2: I don't know if the data is being useful for somebody doing something. It's going through finding the exceptions and then moving beacons.

DO1: [to the interviewer] Do you have an example of the density?

Int: in a certain area, for example, in a park, there should be a certain number of trees.

If your trees are too close they should be classified as an area of trees instead of points.

DO1: What if for example some of our buildings, they need to be within a certain distance within a neighbourhood. That would be a rule about density?

Int: Probably more related with distance.

The thing is that if there is something from a policy that we can check, ok, we want to understand if this is the case or if it is something that should be in a different location.

GIS1: So that would be the quality not of our data but the quality of the world … if we are following the rules in the real world. We can't tell if the data is wrong.

Int: Actually, it is good to have an indicator, so you can check how the real things are fulfilling the law, but that would be an analysis to do more than a check to realize if someone entered a building in the wrong place in the dataset, isn't it?

GIS2: Are you talking about your asset register?

DO1: It is similar but, you know when I need to have a buffer from a certain property... it's because that… It's kind of a demographic thing, for example, there are some commercial assets, and it is supposed to have a certain amount of shops in each ward

DO2: So, mine is about shape: the depiction of them should be the same as the direction of the one-way restriction.

DO1: Another: Within flats there are some ground floor shops and how do we visually see each shop without having to inspect the addresses? Most residential contains shops and where each shop in that flat so you see just the flats itself.

GIS1: How do you imagine it?

DO1: we have no layer with shops yet, again the same problem.

GIS1: If you have polygons with these shops then these polygons should match with the flats polygons, or at least inside the building.

GIS2: That constraints would be for a new layer.

DO2: I have one about direction, I think: the depiction of a one-way street should have arrows on it, in the direction of the one-way street restriction.

DO2: This one is that the LBH managed roads should be wholely within the LBH public highway.

This one is the Transport for London owned roads they should match, should not overlap, and should not have gaps

And this is that the housing assets should be within the housing estates polygons

GIS2: I have two more as well. One is that polygons [in a park] cannot overlap each other unless one of them is a playing area. That's the only one that can overlap another surface. So, you cannot have more than one vegetation in the same one or tarmac and vegetation.

GIS1: And do you have defined that in a park you have a complete partition of the space covered by all the polygons? without gaps between?

GIS2: They shouldn't

GIS1: Ok so that is a partition of the space, so the space is entirely covered by the parts.

GIS2: Actually, it is now, the one that are not clear enough. That could be a rule.

DO1: There are car parks that can only capture the car space and not grass. They should only be pavement.

GIS1: If some object called car park is on grass is it meant that there shouldn't be a carpark there or that the data is wrong.

DO1: The grass should not be covered by the carpark polygon.

Transcription of additional question made to DO2 in personal interview:

Int: about this rule [the interviewer shows the form with the rule "crossing polygon + zebra beacon point should be within x meters of each other and not within"], I couldn't find the data about the zebra crossing polygons.

DO2: Oh yes, it's not in earthlight yet, neither in the server, so maybe forgot about it.

Int: ah, I see, and the same for this one [the interviewer shows the form with the rule "footway + carriageway should not overlap or have gaps between them"], to which layers do you refer?

DO2: Oh, those are not in Earthlight either.

Int: and here in this one [the interviewer shows the form with the rule "Highway assets should only appear within the polygon for public highway"], which are the assets to be considered?

DO2: those either. They are in Mayrise [asset management software] and we haven't migrated everything yet. Earthlight is very new.

Int: and how difficult is to upload them to earthlight?

DO2: Well, it's not that easy [very technical explanation about how to do it]

Int: Well, let me know if you upload some of this during this week but don't worry if you don't.

And [about the same rule] I'm assuming the points and line assets can't be in the boundary.

DO2: They shouldn't be... I think.

Int: Ok. And about this rule "Red route and LBH pub highway should not overlap and not have gaps". Is there one of the layers which should always be the anchor (the unchangeable)?

DO2: The red route but it's tricky. It's because it's the official.

Int: Related to this rule [the interviewer shows the form with the rule "private roads should be wholely outside of public highway"], how you would express it using this terminology which is the standardized one [the interviewer shows the 8 named predicates that ISO suggest using]

DO2: They shouldn't overlap but they can touch.

Int: Ok, and what if I show you these diagrams that explains the relations. There's nothing here [in the overlap row] because the overlap relation is not defined for lines and polygons.

DO2: Oh...so, they shouldn't be crossing.

Int: Perfect, it was just to confirm that they are not that clear and sometimes the word that we commonly use have not the same meaning as in the standards.

I checked the rule and there are 238 out of 276 private roads not crossing or within public highways polygons, an 86%. Is this ok? It is still a rule?

DO2: yes. This data come from a road network layer and that's why most of the road lines are crossing in the public highways, but for me, I need them not crossing.

Int: Ok. And if you trim the [private road] line, what should happen with the trimmed part? Shouldn't be part of the central line of the public highway line, right?

DO2: right, because it's not a network layer

Int: And there are some of them that are not only crossing, they are entirely within the public highway polygon.

DO2: Oh, that's important to know.

Int: would you like to be able to distinguish them as a specific issue?

DO2: Is it possible? Maybe if you calculate the percentage that is crossing in.

Int: yes, and also is possible to check which private roads are within, and that means entirely within.

Right, about this second rule "LBH managed roads should be wholely within LBH public highway", why you think is important to say "wholely" within and not just "within".

DO2: just to be clear, and because it can be understood that when a part is inside it is also within.

Int: Yes. OK, and I wanted to know what happen if an LBH managed road is partly in one LBH public Highway and partly in another

DO2: That's fine.

Int: Perfect. And the same for this one [the interviewer shows the form with the rule "All housing estate assets should be within housing estate polygon"], can a polygonal or linear asset be in the middle of two estates that are touching each other?

DO2: I think so. For me the important is that they don't appear in the polygons we manage.

Int: But you don't have to manage an asset if it is from housing, right?

DO2: not always. Sometimes is not that clear.

Appendix 12: Transcription of the second workshop notes

Participants:

- DO3: Data owner number 3
- DO4: Data owner number 4
- DO5: Data owner number 5
- GIS 1: GIS team member number 1
- GIS2: GIS team member number 2
- Int: Interviewer / author

Notes:

- The squared brackets correspond to comments of the author to try to explain things that were expressed in a non-verbal communication.

DO3: My first one is: Identify all recycling bins not within estates or not within 5 m of estate...to check if this had been capturing correctly or they need to be moved inside.

And this is looking on a different layer created within recycling or waste, and another layer maintained by housing, so they have got an LBH estates and you've got a recycling estate. So, there is two sorts of queries or analysis: one is "identifying where polygons do not overlap" and then "identify polygons which do overlap so where they do overlap or praised (?)" together then we know, ok, we are talking here about the same estate, for the analysis thereafter. But where they don't overlap then there is a slight issue because we are considering in the estates to be a certain area and they considering that the estate to be something different.

GIS1: Who would be the person who be bothered upon that?

DO3: It would be both of us [housing, recycling] because it could be that we have captured incorrectly, or they have captured incorrectly and then, in terms of broader consequences, when costing is done so when work is done under a state and we are costing it or charging housing, then the area is going to make a difference in the work that it's been done, so it relies on charging more or minus…

And as well as within the estates there's public highways which they would do maintaining, etc. so they need to know the area of those things

GIS1: you think the overlap test be enough or it should be …

DO3: it should be almost matching

DO3: Another one: Looking at rec.estates and the LLPG, the address layer identifies which rec.estate do not have a "BLPU property" classification. For us that is what we would need to know is that potentially that state has street level services so if it states are blocks of flats but also if they are individual terraced house or semi-detached, which we as a waste & recycling supply to different type of services so we would need to know that to ensure that we are giving them the correct service.

DO4: My use of data is slightly different because I'm accessing data from different places but I'm trying to make a list of the source of data access we use, so points data would be health services so (GP practitioners, and that sort of things) so we don't particularly use the data on Earthlight so for example a dentistry services should be within borough boundaries.

Generally, our degree of accuracy would be something that looks ok on a map rather that literally something that we could show to commissioners and that sort of things. Sometimes there are issues where we need to separate things that are in the same location manually to show them more clearly in the map.

My rules are:

First, point data, such as general practitioners, opticians, pharmacies, hospitals, dentists, clinics, should be within borough boundaries. But we need flexibility. Sometimes two practitioners are in the same building and we need to move the points manually in the map to make clear that there are two practitioners there.

Second, the boundaries (LSOAS, wards, grouped LSOAS and neighbourhoods should be reasonable accurate for visual display and for calculating density, rates per ward, travel times

Int: But what would you say is reasonable accurate?

DO4: just close to the real location, and normally our data is external, and it is correct, more or less, but for the scale we do the maps we don´t need something very accurate

The next one is the same but for other locations related with health determinants, such as fast food premises, leisure centres, parks, pubs, betting shops, etc.

DO5: There are nationally listed buildings and locally listed buildings and a building shouldn't be both.

DO5: This is about shapes when I'm doing listed buildings it could be nationally or locally listed the shape that I'm copying and pasting normally, it should be the same as the one on the OSMM because some of the data results, when you put on the layer you put building extensions and things and the building listed is slightly different shape from what is now in the OS.

GIS1: That could be as a trigger to know where to update the data

DO5: There is a kind of conceptual curtilage where if you list a building, legally, things in the garden and the walls and railings and boundaries are on the curtilage and they are listed as well, so in theory requires not to show that. The problem is that is quite complicated because there are extra rules: if they fall in the curtilage and they are built after 1948 they are not included and sometimes the idea is what is the curtilage of a building can be quite complicated because it might have a big garden and then sometimes the garden might be sold off…so it can be a very complicate decision when people ask me to decide based in the law, but we can think about shrunk curtilage. But sometimes it's really clear but other ones are really not. We have a woman that owns a house from the 1680 which originally had a garden but at one point the garden was turned into a 19's century workshop on the back so it was two buildings. Before she bought them the two buildings were in one ownership and the whole house was the office of the workshop so at the point it was listed the workshop was part of the same curtilage but the curtilage was then subdivided after the listing and so she said is that building in the back listed or not? Because the neighbour wants to down it to build a block of flats which she doesn't want to…

GIS1: do you see a way to define that?

DO5: No, I don't know

GIS1: it seems that you can't define a rule...

DO5: no

DO5: [Another one] We also designate locally listed (LL) buildings and we designated some LL buildings in the area that is kind of out of the control of the London legacy development corporation, they are the planning authority. So those LL buildings are actually out of ours and now the developers question it because we designated something that was not under our control. It's in the borough of hackney but where hackney is no the planning authority at the moment because is the legacy of the Olympics.

The case is complicate because we have a listed school in that area, and they are asking me and I'm saying I'm not the planning authority even the schools belongs to hackney and it's run by hackney's education service and hackney is the landowner we are not the local planning authority, so they didn't have to deal with the conditions …

GIS1: Maybe this object needs a field that says that this building is not hackneys responsibility.

DO5: That would work

DO5: [Another one:] A conservation area boundary should not cut a building in half. The clever thing is to define the boundary in the street or over a wall but shouldn't go through a building. I mean, it can be but is not really the idea.

DO5: [Another one:] A building should not be in more than one conservation area. And that happens. There is a building that is in two. Another conservation area, a narrow one which the regents canal east west and another one is Kingsland road which goes north south and there is a building in the middle. Apparently is legal but is not the idea.

DO5: [Another one:] A Nationally or locally listed building should include data on: asset type (e.g. building or street furniture), ownership (e.g. council or private).

DO5: [Another one:] The shape of a building should be the shape on OSMM (with exceptions)

DO5: [Another one:] Hackney's data on nationally listed buildings does not exactly match Historic England data (and it should)

Int: You didn't draw anything because of time or because you think is better explaining it.

DO5: Oh! I just didn't think about that

DO5: The thing about the relationship between the listed building and the OS is funny because generally the listed buildings odd to be as on OS but there's a case where there is a listed building which isn't on the OS map because it'd been removed. And then there is one building where the owning listed is part of a larger building so there's actually lying within the building where some of it its listed and some it's not, which was a really silly decision but that's what they did.

GIS1: That's why would be a report that show little images so if you think it's an exception you can manage.

So, you can't have listed buildings that are half of a building, but can you have some that are a building plus a garden, or plus, you know, there are several master map polygons matched together…

DO5: yeah. So, it's a complex rule because how many cases…

Int: [general final question:] How do you think the definition of constraints can be eased for you…

DO3: Asking for how we don't want the data to be. That's a constraint. It was confusing with analysis, that's why I didn't write rules, I wrote the type of analysis I would need to do to check if there are problems.

GIS1: It would be more direct

GIS2 so instead of "how do you want your data to look like it's how it shouldn't look like.

DO5: My concern is if you are going to create GIS rules to prevent data entry based on these rules, because there are some situations that have exceptions. If the system becomes too rigid, it is complicated. I worked previously in a planning database in another council where were rules to stop certain data entries and you ended with the most load work trying to cheat the system, using redundant fields to explain things and it was really scrappy and painful, so I think it should not be many rules.

GIS1: would be that we check but we don't prevent

Transcription of additional question made to DO2 in personal interview:

Int: From the form A n1 ["a rec.bin should not be further than 5 m from an estate"], which is the rule that we can extract? Or, what you don't want to see?

DO3: A recycling estate polygon without a point within it or a point within a certain distance

Int: but it can happen that a recycling bin is less than 5 m from more than one estate, so you would not be able to define to which estate it corresponds [The interviewer shows situations like that]

DO3: Maybe we just say that the recycling bins should be within the recycling estates. And that means that the recycling estate should include more than the building footprint. It is not clear how to define a rule to determine which polygons of the OS mastermap the recycling estates should follow, so let's don't define it yet, at least for this stage. It should not only consider the building footprint but also the access road, but I'm not sure about that… or maybe the "surrounding ground". The important thing is that a recycling bin should be within it, though there are some exceptions. It would be helpful if we can define exceptions to the query, so they don't come up again in the report.

Int: And related to the second form, where you put "identify where polygons [LBH estates and REC estates] do not overlap and where polygons overlap

DO3: It is related with the same, trying to define something that should be followed by the recycling estates. Maybe we can just say that even if they don't have to be coincident, the recycling estates should snap to the OS mastermap lines

Int: And this one. I didn´t understand what is the BLPU

DO3: It is the use and when the BLPU is RD06 it means that it is a flat. So, addresses with BLPU different to RD06 should not be within recycling estates.

Transcription of additional question made to DO5 in personal interview:

Int: In one of your rules you said: "Hackney's data on nationally listed buildings does not exactly match Historic England data, and it should". Is the Historic England data in a layer in Earthlight? where I can find it?

DO5: The layer is in the M drive under Planning and there are layers for statutorily listed buildings, locally listed buildings and Conservation Areas.

Int: And, just to confirm, the nationally listed buildings correspond to the "statutory listed buildings", right?

DO5: They are the same thing, yes.

Int: An in this one you said: "A conservation area should not cut a building in half". What happened if it cut a building but not in a half, or even in a small part. Is still a problem? Or you mean that the problem is when the building is cut in an important proportion

DO5: No, a conservation area should not cut a building at all. It should not say "in half" it may lead to misunderstanding.

Appendix 13: Coverage of the topological relations by the spatial constraints expressed

	Disjoint	Equals	Touches	Crosses	Overlaps	Contains	Within
Point / point	X	X				X	X
Point / line	X		X				X
Line / Point						X	
Point / poly	X		X				✓
Poly / point						X	
Line / Line	X	X	X	X	X	X	X
Line / Poly	X		X	✓			✓
Poly / Line						X	
Poly / Poly	X	✓	X		✓	X	✓

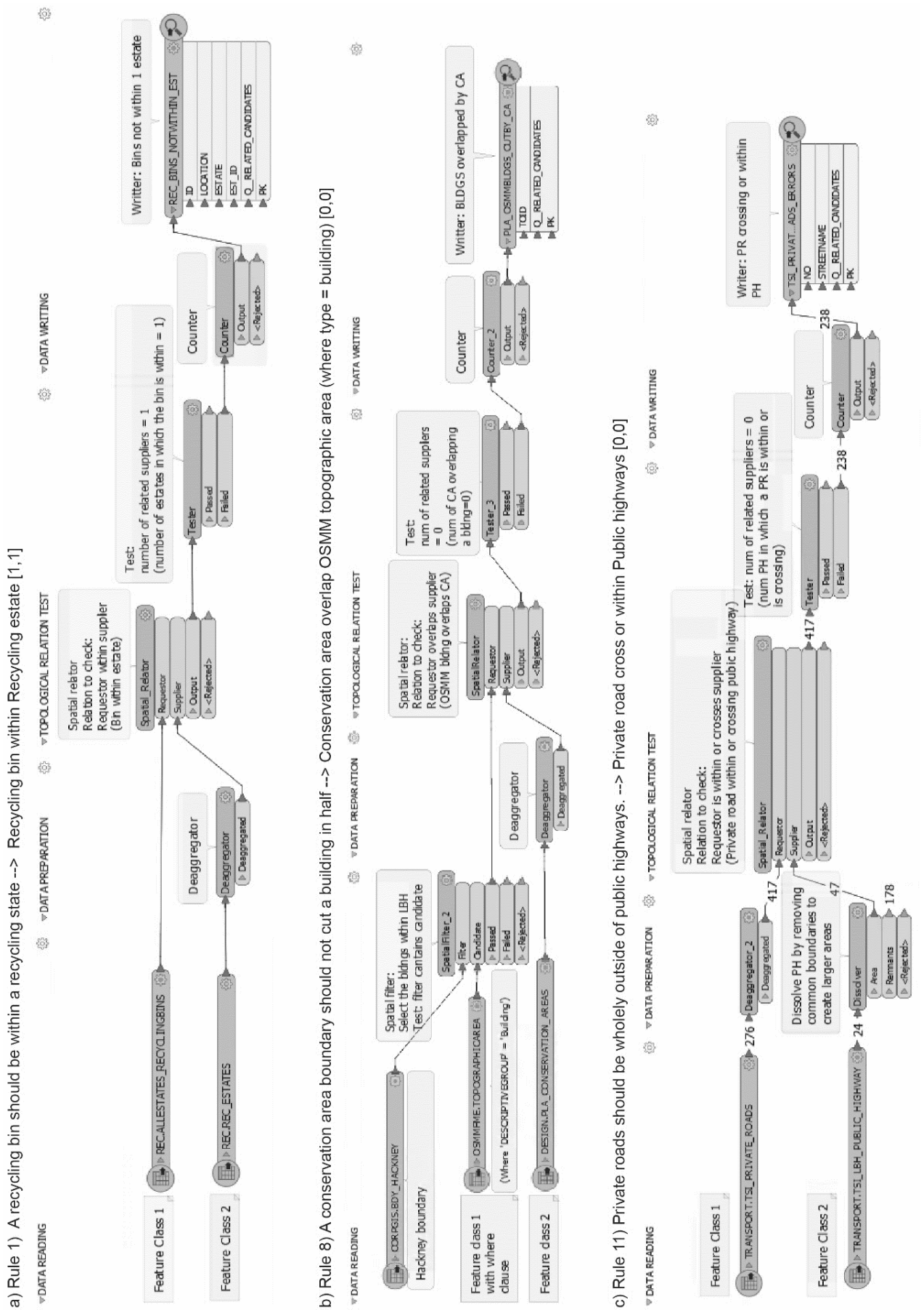
The colour grey indicates that the relation is geometrically impossible according to the 9IM
The ✓ indicates that there were constraints expressed which implies that topological relation.
The X indicates that there were not constraints expressed which implies that topological relation.

Appendix 14: Result of the translation of the spatial relations implicit in the constraints

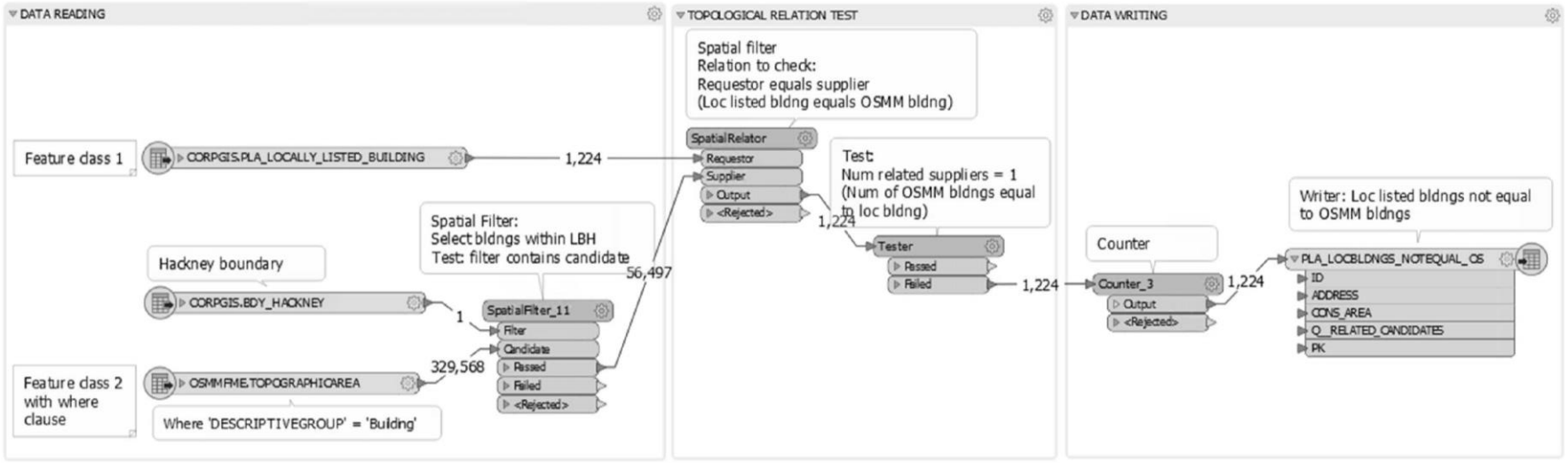
	Disjoint	Equals	Touches	Crosses	Overlaps	Contains	Within
Point / Polygon							
REC_1) A recycling bin should be within a recycling estate.	F		F				C
REC_2) Addresses with BLPU different to RD06 (flats) should not be within recycling states *	C		F				F
HEAL_1) Point data (general practitioners / opticians / pharmacies / hospitals / dentists / clinics) [should be] within borough boundaries.	F		F				C
Line / Polygon							
STRE_1) Private roads should be wholely outside of public highways.	A		A	F			F
STRE_3) LBH managed road wholely within [the union of] LBH public highway	F		F	F			C
STRE_2a) [Linear] Highway assets should only appear within [the union of] public highway	F		F	F			C
Polygon / Polygon							
CORP_2) Corporate building with no extension into the road	A	F	A		F	F	F
CORP_3) Car parks areas should not include any topographic area covered with grass.	A	F	A		F	F	F
CONS_1) A conservation area must be in Hackney	F	F	F		F	F	C
CONS_2) A building should not be in more than one conservation area	A	F	A		F	F	F[2,n]
CONS_3) A conservation area boundary should not cut a building	A	F	A		F	A	F
CONS 5) Hackney's data on nationally listed buildings does not exactly match Historic England data (and it should)	A	C	A		F	F	F
CONS 6) The shape of a listed building should be the shape on OSMM (with exceptions)	A	C	A		F	F	F
STRE_2b) [Polygonal] Highway assets should only appear within the [the union of] public highway	F	F	F		F	F	C

The colour grey indicates that the relation is geometrically impossible according to the 9IM
"A" indicates that the relation is allowed, C indicates that it is compulsory, F indicates that it is forbidden.
When not specified, the forbidden relations have [0,0] cardinality and the compulsory [1,n]

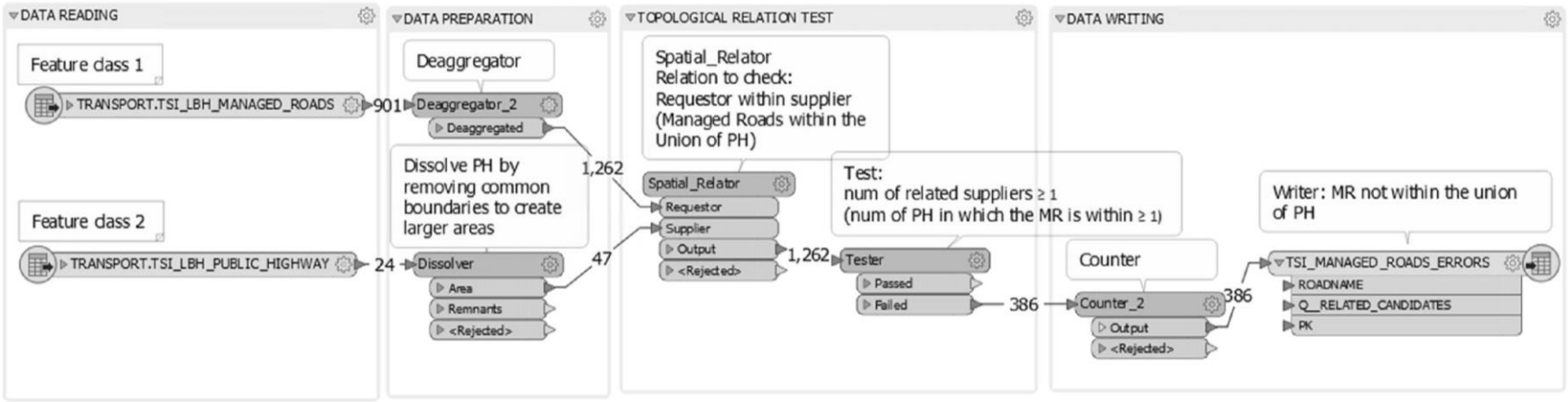
Appendix 15: Basic checks workspaces



a) Rule 10b) The shape of a locally listed building should be the dhape on OSMM --> Locally listed building equal OSMM topographic area (where type = building) [1,1]



b) Rule 14) LBH managed road wholely within LBH public highway --> LBH managed road within (the union of) Public highways [1,1]

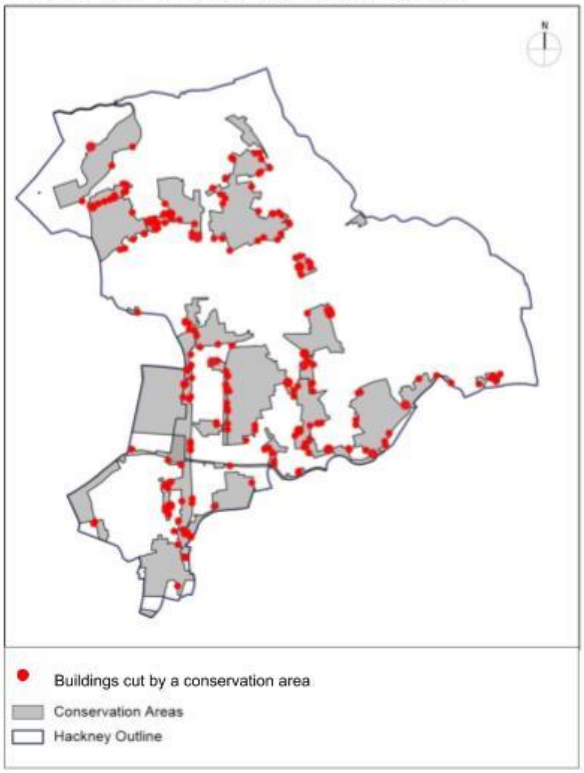


Appendix 16: Basic reports

a) Report for rule “a conservation area should not cut a building”

Data quality checking report

“A conservation area should not cut a building”



The conservation areas are cutting

245

buildings in Hackney.

Where to find the data: *(link to location or to download)*

Examples:

TOID	Conservation area	CA_ID
5000005170592835	Clapton Square	1
5000005186494818	Graham Road and Mapledene	18

TOID	Conservation area	CA_ID
1000006316847	Clapton Square	1
1000006042423	Dalston	31

TOID	Conservation area	CA_ID
1000006042402	Dalston	31
1000001802845624	Dalston lane (west)	24

Alternatives to solve these issues:

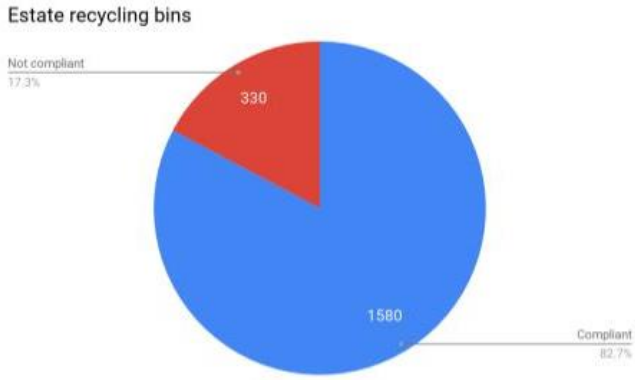
- Manually correct the conservation area polygon to eliminate overlaps.
- *Mark the building as an exception* (link)

b) Report for rule “a recycling bin should be within a recycling estate”

Data quality checking report

"An estate recycling bin should be within a recycling estate"

230 out of 1910 estate recycling bins are not within one (and only one) recycling estate.



Where to find the data: *(link to location or to download)*

Examples:

ID	ESTATE	EST_ID
998	The Limes - 5 Massie rd	588
2050	Armstrong house, 146, Southwold Road	51
289	Armstrong house, 146, Southwold Road	51

ID	ESTATE	EST_ID
998	The Limes - 5 Massie rd	588
2050	Armstrong house, 146, Southwold Road	51
289	Armstrong house, 146, Southwold Road	51

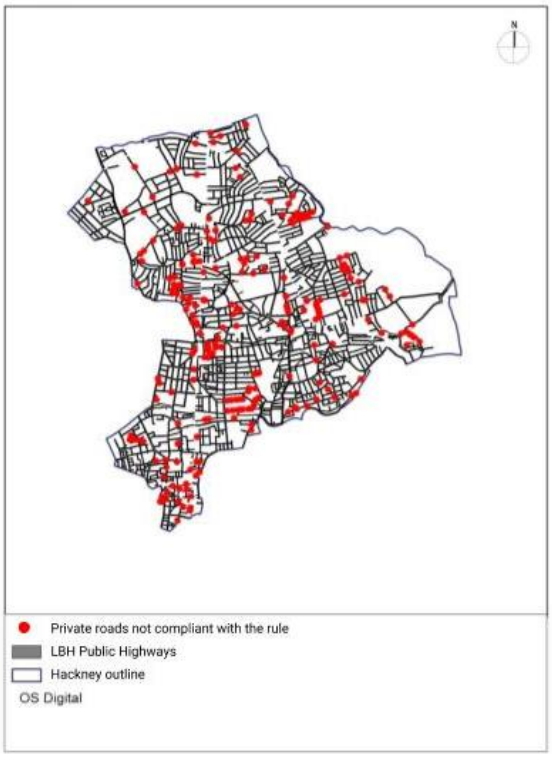
Alternatives to solve these issues:

- Manually correct the recycling estate polygon to make it contain the recycling bin.
- Manually correct the position of the recycling bin.
- *Identify the bin as an exception.* (link)

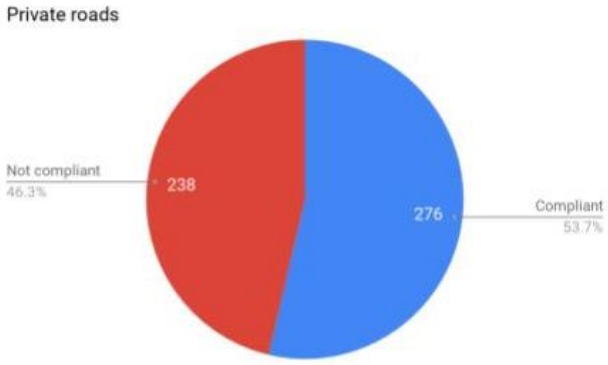
c) Report for rule “private roads should be wholely outside public highway”

Data quality checking report

"Private roads should be wholely outside of public highway"



238 out of 276 private roads are not wholely outside of public highway.

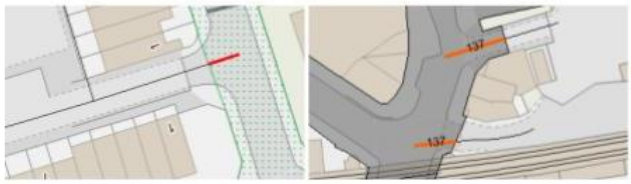


Note: the total number of private roads here corresponds to its number after deaggregating.

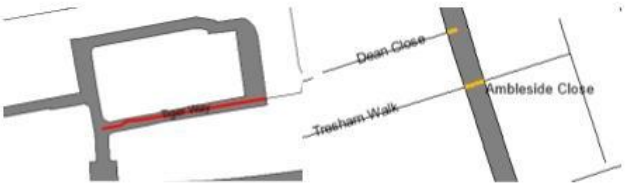
Where to find the data: *(link to location or to download)*

Examples:

STREET NAME	Length inside public highway	NO
Lyme Grove	80.4595m	673
Cambridge passage	38.8389 m	196



STREET NAME	Length inside public highway	NO
Burnett Close	5.5487 m	193
Bohemia place	18.9632 m	137
Bohemia place	12.4394 m	137

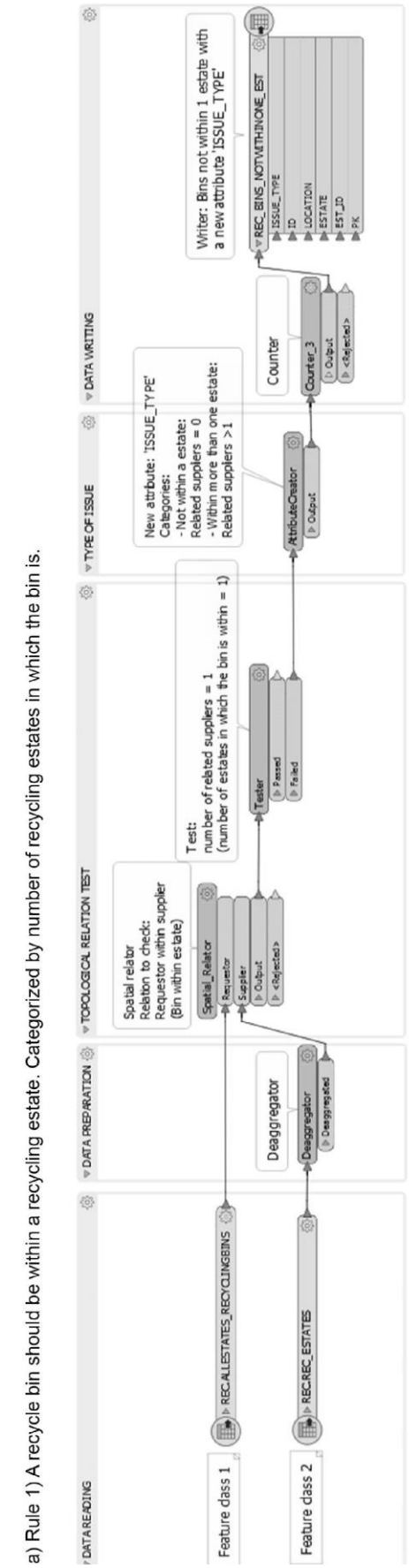


STREET NAME	Length inside public highway	NO
Tiger way	126.3782 m	1064
Tresham Walk	5.3608 m	1075
Ambleside Close	5.0061 m	34

Alternatives to solve these issues:

- Manually correct the private road to make it compliant.
- *Identify the private road as an exception.* (link)

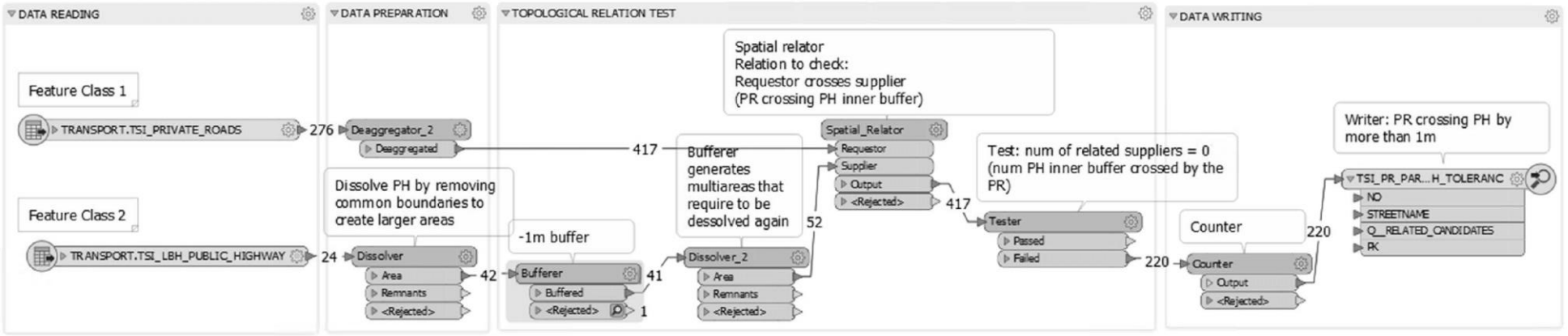
Appendix 17: Detailed checks workspaces



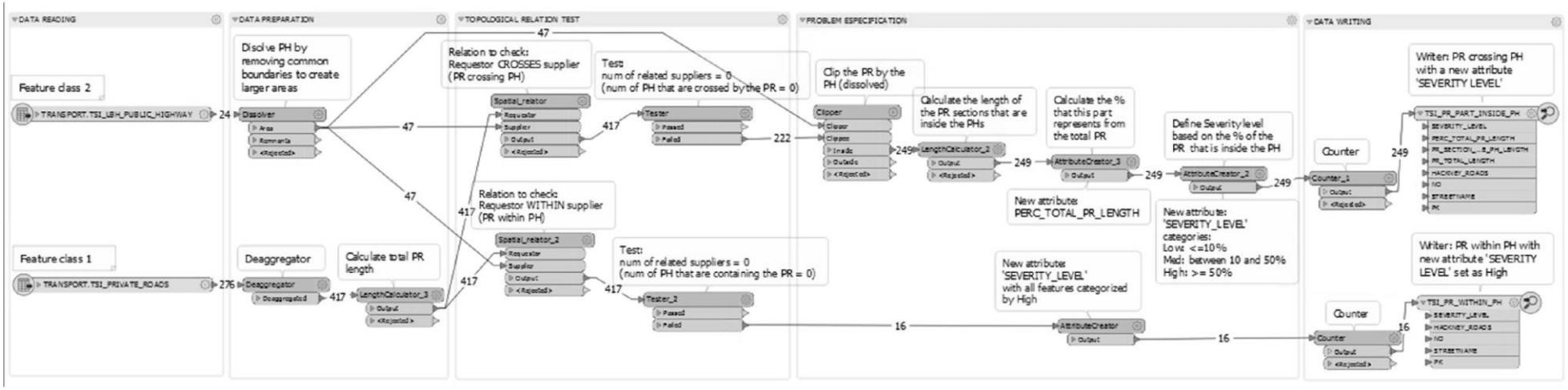
a) Rule 1) A recycle bin should be within a recycling estate. Categorized by number of recycling estates in which the bin is.

b) Rule 10b) The shape of a locally listed building should be the shape on OSMM. Buffer approach

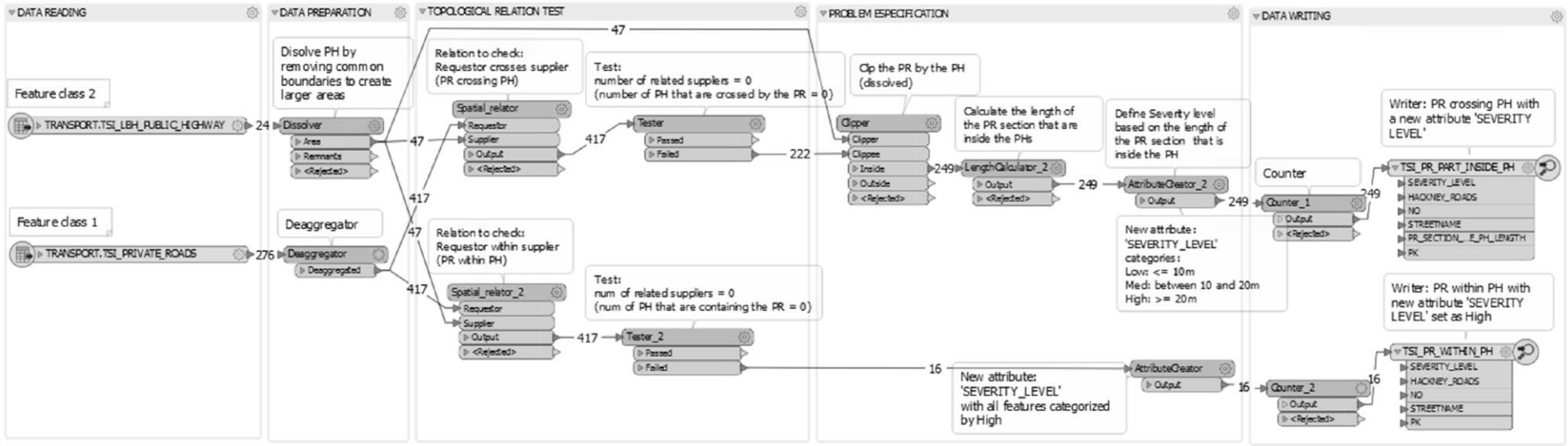
c.1) Buffer approach Rule 11) “Private road should be wholely outside public highways” --> Private road cross or within Public highway [0,0]



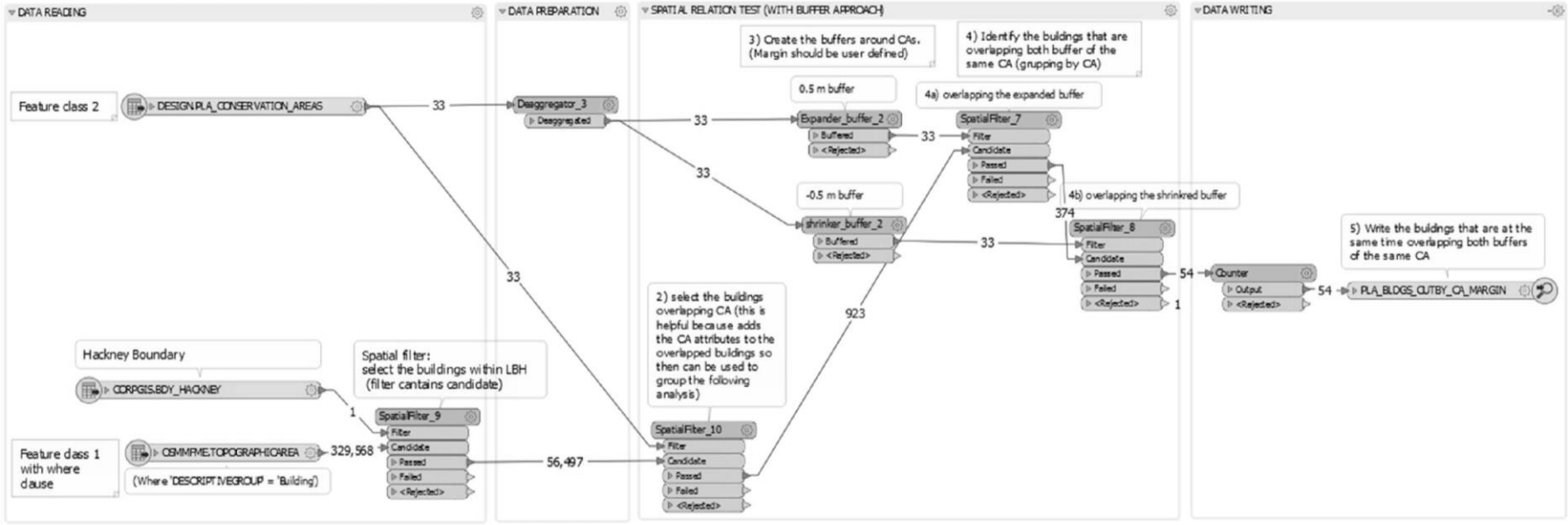
c.2) Percentage approach Rule 11)



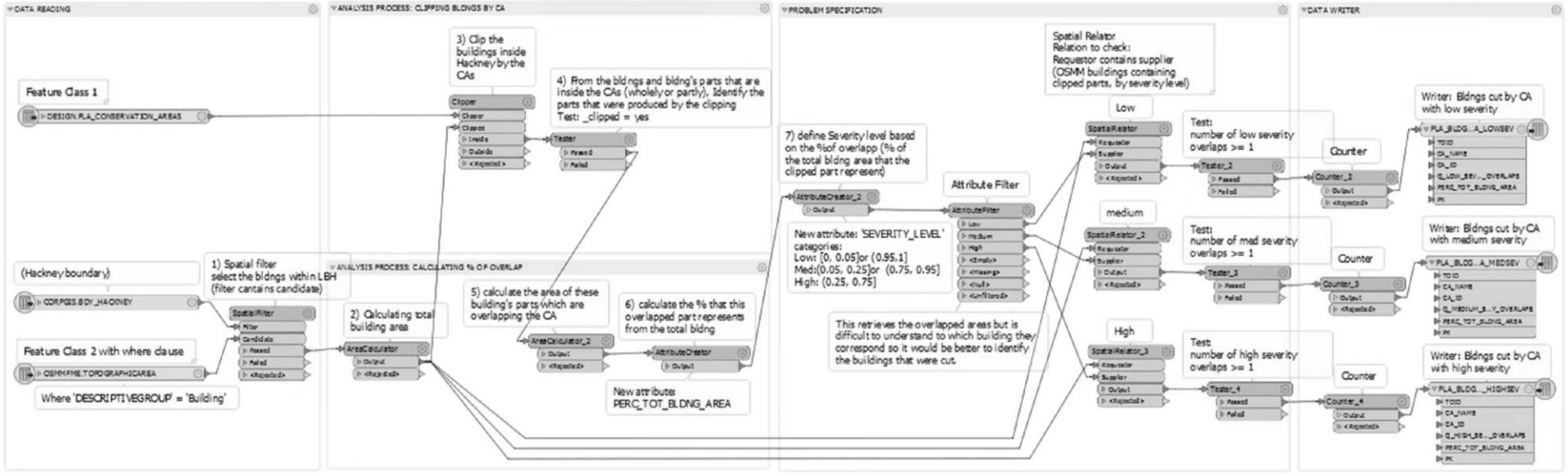
c.3) Length approach Rule 11) “Private road should be wholely outside public highways” --> Private road cross or within Public highway [0,0]



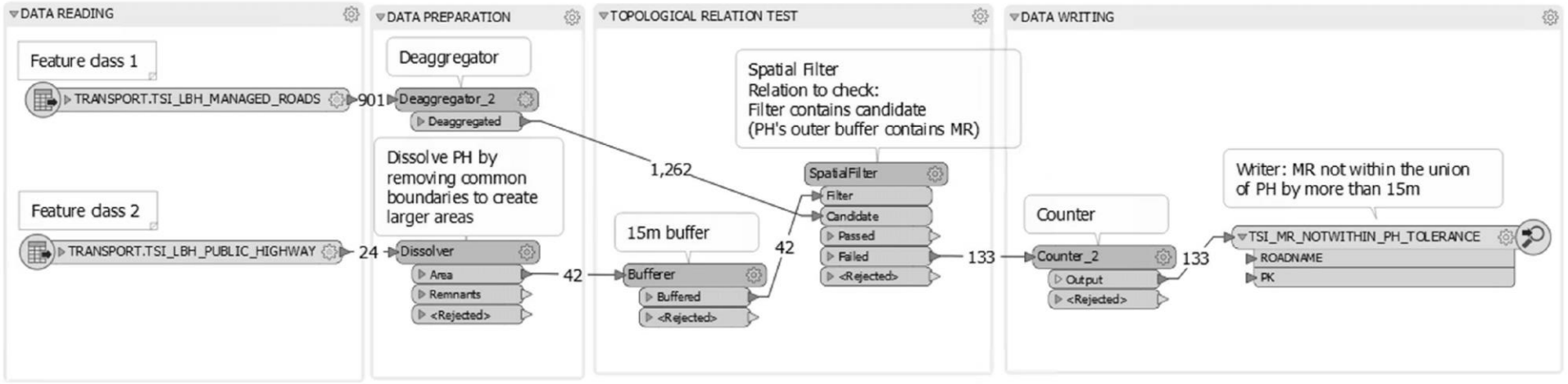
d.1) Buffer approach Rule 8) “A conservation area should not cut a building”



d.2) Percentage approach for Rule 8



e) Buffer approach Rule 14) “LBH managed roads should be wholely within LBH public highway”

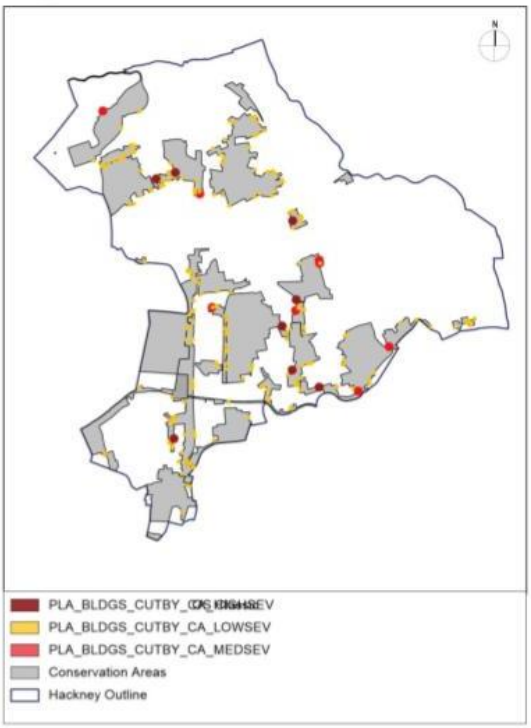


Appendix 18: Detailed reports

a) Report for rule "a conservation area should not cut a building"

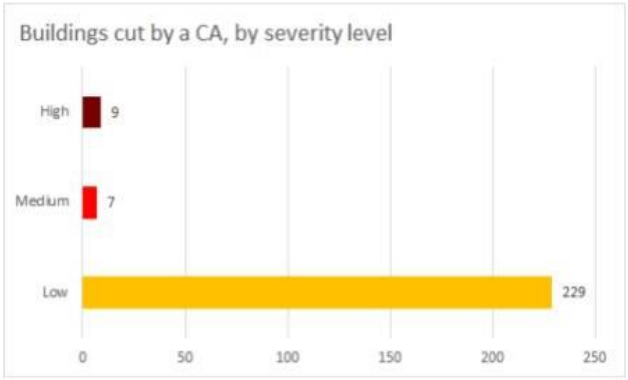
Data quality checking report

"A conservation area should not cut a building"



The conservation areas are cutting 245 buildings in Hackney, 9 of them with high severity, 7 with medium severity and the rest with low severity (see criteria for severity level at the bottom).

Notice that a building can be cut in more than one part so a building can have issues with different severity levels.



Where to find the data: *(link to location or to download)*

Examples:

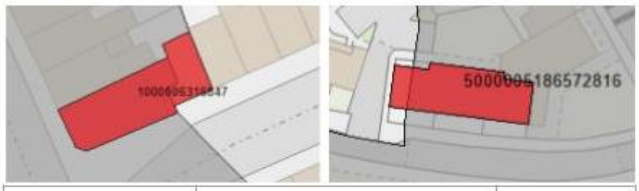
Situation 1: High severity level

(25% - 75% of the building is inside the CA)

TOID	Conservation area	CA_ID
5000005170592835	Clapton Square	1
5000005186494818	Graham Road and Mapledene	18

Situation 2: Medium severity level

(5% - 25% or 75% - 95% of the building is inside the CA)



TOID	Conservation area	CA_ID
1000006316847	Clapton Square	1
5000005186572816	Victoria Park	6

Situation 3: Low severity level

(Less than 5% or more than 95% is inside the CA)

TOID	Conservation area	CA_ID
5000005151934514	Stoke Newington Reservoirs, Filter Beds and New River	11
1000006042423	Dalston	31

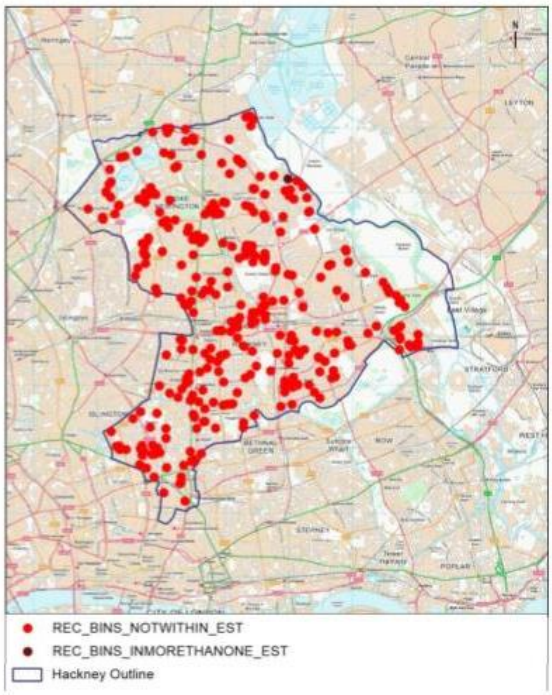
Alternatives to solve these issues:

- Manually correct the conservation area polygon to eliminate overlaps.
- *Mark the building as an exception* (link)

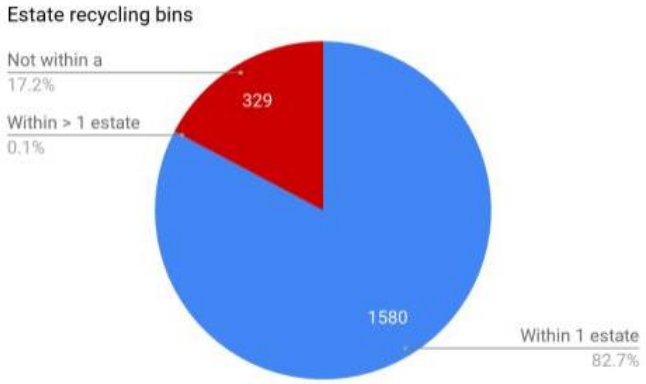
a) Report for rule “a recycling bin should be within a recycling estate”

Data quality checking report

"An estate recycling bin should be within a recycling estate"



330 out of 1,910 bins are not compliant with the rule (17.3%). 329 are not within a estate (situation 1) and 1 are within more than one estate (situation 2).



Problematic situations:	Number	Percentage
Bins not within a estate	329	17.2%
Bins within more than one estate	1	0.1%

Where to find the data: *(link to location or to download)*

Examples:

Situation 1: Bins not within a estate

ID	ESTATE	EST_ID
998	The Limes - 5 Massie rd	588
2050	Armstrong house, 146, Southwold Road	51
289	Armstrong house, 146, Southwold Road	51

Alternatives to solve these issues:

- Manually correct the recycling estate polygon to make it contain the recycling bin.
- Manually correct the position of the recycling bin.
- *Identify the bin as an exception.* (link)

Situation 2: Bins within more than 1 estate

ID	ESTATE	EST_ID
806	Courtlands	817

Alternatives to solve these issues:

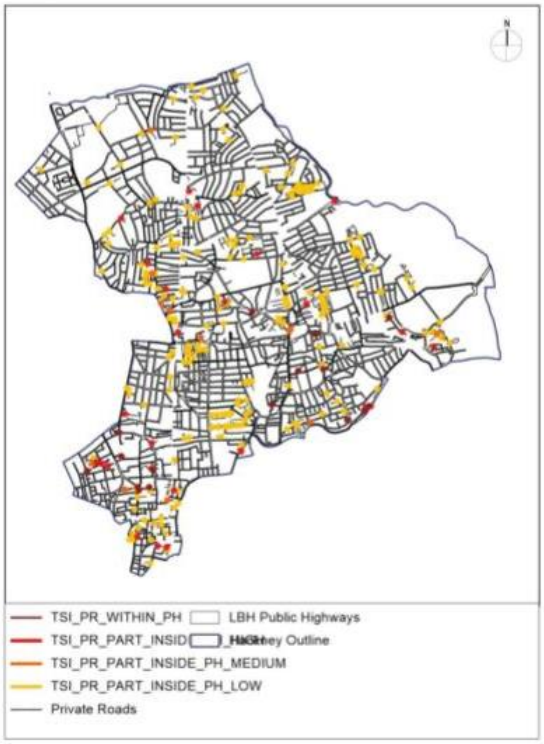
- Manually correct the recycling estate polygons to eliminate overlaps.

Where to find the data: *(link to location or to download)*

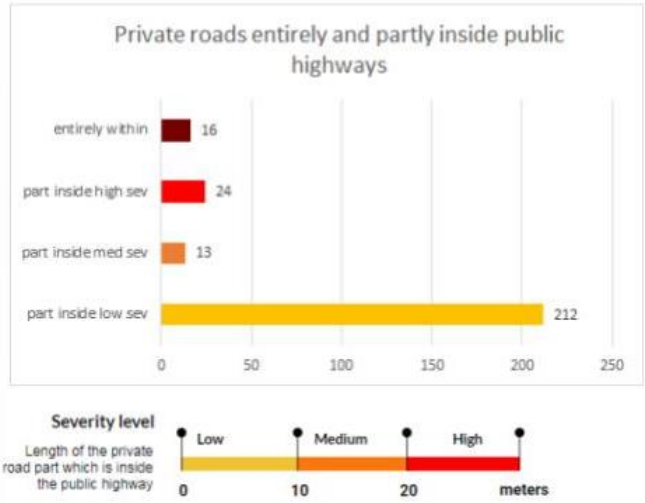
b) Report for rule “a private road should be wholely outside of public highway”

Data quality checking report

“Private roads should be wholely outside of public highway”



16 private roads (PR) are entirely within public highways (PH) and 222 are partly inside them. Distinguishing the latter by severity level based on the length of the PR part that is inside the PH , there are 24 high severity issues, 23 medium severity issues and 212 low severity issues. Notice that a PR can have more than one part inside PH with different severity levels.



Where to find the data: *(link to location or to download)*

Examples:

Situation 1: Private road entirely within public highway

STREET NAME	Length	NO
Lyme Grove	80.4595m	673
Cambridge passage	38.8389 m	196

Alternatives to solve these issues:

- ?????
- *Mark the private road as an exception* (link)

Situation 2: High severity level

(A section bigger than 20m is inside the PH)

STREET NAME	Length	NO
Tiger way	126.3782 m	1064
Gibson Gardens	24.2650 m	451

Alternatives to solve these issues:

- ??????
- *Mark the building as an exception* (link)

Examples:

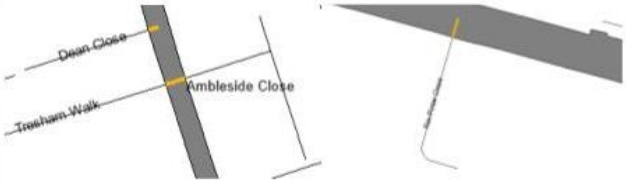
Situation 3: Medium severity level
(A section with length between 10m and 20m is inside the PH)

Street name	Length inside (meters)	NO
Bohemia place	12.4394	137
Tavistock Close	13.2250	1053

Alternatives to solve these issues:

- Delete the part of the private road which is inside the PH polygon
- *Mark the private road as an exception* (link)

Situation 4: Low severity level
(A section with length 10m or less is inside the PH)



Street name	Length inside (meters)	NO
Dean Close	5.3263 m	328
Tresham Walk	5.3608 m	1075
Ambleside Close	5.0061 m	34
Rav Pinter Close	9.9144 m	882

Alternatives to solve these issues:

- Delete the part of the private road which is inside the PH polygon
- *Mark the private road as an exception* (link)

Appendix 19: Report feedback interview. Data owner n°2

REPORT FEEDBACK FROM DO2
July 25

Questions about basic reports:

Int: In general, what do you think about a report like this:

DO2: It's good. I can see what the issues are. It's good to know how many there are and the relation with the total. The general map shows the problem is across all the area and not on isolated spots. Examples are good.

Int: How useful is for you as a tool to improve the quality of your data?

DO2: It's useful because it raises a problem

Int: What you would do with this data?

DO2: I would change them. The report tells you that there is an issue. And if I change something I can check if it was done fine.

Int: and in concrete, how you think you may correct these issues

DO2: I would go to each of them to see what to do. Some of them may be completely wrong it might not be a private road at all. 238 isn't that many. But, if it's possible it would help to create an automated process to correct them.

Int: What do you think about the chart type, a pie chart?
DO2: its fine.

Int: The total number represents the number of private roads after disaggregating them. Is that fine?
DO2: Yes, if it is explained.

Int: In the 2nd page, the examples are shown using different styles. What do you think about the following aspects of the maps: Presence of base map?

DO2: is better to have it, as reference.

Int: Presence of cartographic text in the base map?

DO2: it is helpful to recognize the places.

Int: Reference layers colour?

DO2:I prefer the colour currently in use in Earthlight. But it depends, it doesn't matter. The issue should be obvious. It can be changed if the original colour of the layer makes the issue not obvious.

Int: Colour and size of the features being checked?

DO2: it doesn't matter, any colour that make it obvious

Int: Highlighting only the part of the private road that is inside a public highway?

DO2: that's good. I prefer that, only the issue highlighted.

Int: Labels?
I don't need a label as far as the carto text is there. The only thing I may need is the street name if it doesn't appear in the carto text, I'm going to go to the map anyway.

Int: What you would like to change to the report to make it more useful or easy to understand?

DO2: nothing

Int: Apart from the report, how you would like to receive the data of the features with issues?

DO2: a tab file, I guess.

Int: Is the data inside the tables the one that you would want to see there? Do you need something else or it´s too much information?

DO2: I just need the street name. If you are doing analysis with the length, maybe differentiating them by the % it would be useful, but if not, I don't need it.

Int: What do you think about the alternatives suggested to solve the issue?

DO2: seems right

Questions about detailed reports:

(Give them a time to check the report alone, without explanation)

Int: Do you think that a more detailed report like this improves the usefulness of the report as a tool to improve the quality of your data?

DO2: Yes, because it's broken down into sections, so it is showing me where to focus, so I'll concentrate on the ones that are entirely within and the ones with high severity, only 40 issues. Where the PR is entirely within probably they are wrong, so this report is good highlighting those cases

Int: What do you think about the idea of defining a "severity level"? And about the name we chose itself.

DO2: That is better than the standard "issue" or "not issue" approach.

Int: Do you agree with the criteria used to define the severity level?

DO2: that´s good.

Int: To define it, I tried to identify types of issues and the low severity corresponds to the length range that typically happens as the private roads layer comes from a network layer.

DO2: It´s more or less the road width right [of the public highway]?

Int: Exactly, the biggest width.

Int: What you would like to change to the report to make it more useful or easy to understand?

DO2: nothing. It´s fine

Int: And what do you think about the chart type which is different?
DO2: this is better because it shows different situations.

Int: But you could cut the pie in more pieces.

DO2: Maybe is a personal thing, I prefer bar charts.

Int: And that it doesn't have the percentages? Because in this case as a private road can have more than one part inside PH with different severity levels, the chart represents the issues and not the private roads.

DO2: That´s fine. The numbers are the important and then I´m going to check them in the data

Int: Do you think the graphic way to express the criteria used to define the severity level is clear and easy to understand?

DO2: yes

Int: In this case, how you would like to receive the data of the features with issues, as 1-tab file for all the features with issues with a severity level attribute or 1-tab file per each severity level?
DO2: 1 tab file with a new attribute

Int: What do you think about the alternatives suggested to solve the issue? And I didn´t know what to suggest for the private roads within public highways

DO2: For that one I would say "Investigate why", and the same for the high severity. Presumably all the low severity can be because of the same reason and can be treated automatically. For the medium severity you have to manually check what to do, but its few of them

Appendix 20: Report feedback interview. Data owner n°3

REPORT FEEDBACK FROM DO3
July 27

Questions about basic reports:

Int: In general, what do you think about a report like this:

DO3: This would be a very good starting point to identify which states we need to review and it raises the issue of the inconsistency on how the states are captured. This one, for example [building at the right in the fourth image from left to right in the 2nd page of the report] just captured the building footprint while others generally include the internal roads. But recently I have been told that there are states where the bin is not within.

Int: Do you think that are exceptions or are more than exceptions?

DO3: I can't say really. I think they are still exceptions.

Int: And about the style, and the examples?

DO3: I prefer the report to be in the way the layer is captured [referring to the colour selected to represent the recycling estates, as in the 3rd and 4th image from left to right in the 2nd page of the report], because is what the end user is more familiar with.

Int: Who do you mean by end user in this case?

DO3: Anyone in my team, the estates recycling officers mainly.

Int: How useful is the report for you as a tool to improve the quality of your data?

DO3: For an initial analysis it is perfect and that would start to improve our data quality

Int: And concretely, what you would do with this data?

DO3: I would look at it individually and compare the recycling estates with the master map to identify what the problem is. But this would be done in conjunction with the states recycling officers. If they are unsure of the location of the bin, a site visit should be done. We will work together to correct the data, so they are going to be clearer on how the data should be captured. What we don´t want is non-compliant records coming in, so the idea is to give the recycling team ownership on maintaining this layer. So, when they capture they would know what the constraints are.

Int: What do you think about the general map:

DO3: it's a good representation because it says it's an issue across all the area, so we can start, for example going from the top to the bottom, and in the future, we would see which states have more problems.

Int: What do you think about not showing the correct ones

DO3. It's fine that the good ones are not, they would clutter all, we want to see the problem areas not everything.

Int: What do you think about the chart type, a pie chart?

DO3: I think the pie chart is easier to read and to see how much work needs to be done, it's easy to understand.

Int: In the 2nd page, the examples are shown using different styles. What do you think about the following aspects of the maps:

Presence of base map?

DO3: That's perfect because we immediately can identify where the issue is

Int: Presence of cartographic text in the base map:

DO3: I prefer to have it to know the location immediately.

Int: About the reference layers colour. You said you preferred in the way the data is in Earthlight. But this are red. Do you think that it can be confusing, like identifying an error?

DO3: Even if it is red, I think in this instance, it should maintain the Earthlight style, because the recycling officer is familiar with the style of the layer, putting anything else is going to be confusing. And in a sense, they are with issues. The issue is not only a problem with the bin location. It can be an error on how the recycling estate was captured.

Int: Colour and size of the features being checked:

DO3: It definitely needs a legend here because when the officer sees a blue dot they think in the blue recycling (food recycling). It's good to have here also the compliant ones because if not they are going to be confused, like “it should be another bin here”, but definitely with a legend.

I would like to have only the estates that are corresponding to the not compliant bins. If it is the same estate with two poly we need to show it but if not, we don't need it. If there are two poly we should be sure that its only one state.

Int: And about the labels?

DO3: For location purposes the bin id is efficient, and the cartographic text to identify the estate

Int: Is the data inside the tables the one that you would want to see there?

DO3: I think for the purpose of the analysis and the info required to correct the error its efficient with that. You don’t want too much info that maybe be confusing. Just the info to highlight the issue.

Int: What you would like to change to the report to make it more useful or easy to understand?

DO3: like we discuss: a legend

Int: What do you think about the alternatives suggested to solve the issue?

DO3: This are actually not alternatives, this is the actual solution. Methods to resolve the issue, instead. Because an alternative is something else.

Int: Apart from the report, how you would like to receive the data of the features with issues?

DO3: a pdf showing this. If we got 330 issues I would have 330 pages with all these issues. So, I can check all of them in the pdf. When we get more into how to solve it I would say a tab file, because you can bring it into MapInfo to join it into rec.bin layers. But seeing all of them in the pdf first we can check what we need to do and define how long it would take to solve the issues. It's sort of time management. It should be included into the day to day work.

Questions about detailed reports:

(Give them a time to check the report alone, without explanation)

Int: Do you think that a more detailed report like this improves the usefulness of the report as a tool to improve the quality of your data?

DO3: With this one the general map, just don't put the blue dots in the legend. Whenever you have a legend, that's what you are showing on the map. The colour for the bin within more than 1 estate is not very contrasting and coherent with the pie chart.

Int: What do you think about the pie chart in this case?

DO3: That’s enough for the purpose. The text, we don't really need it. The pie chart and the table speak for itself. The only thing that is not in the chart but in the text is the total, so It would be nice to have the total number of bins somewhere in the chart. For me, I don't really read too much. Maybe put the text at the bottom, with less importance.

Int: And in terms of the utility of the report?

DO3: The problem is that there’s only one bin within two recycling estates. If it would be a higher percentage that would be more important to solve. But definitely is good to have that analysis and continuing doing it because it highlights the captures inaccuracies and a different type or error.

Int: And what do you think about the alternatives or methods offered this time :

DO3: That's fine. Or maybe remove the duplicate. Delete it if it’s a duplicate polygon.
For this specific situation it would be nice to have both polygons within the bounding box and with different colours. And I would like to see the id of both estates that are containing the bin.

Int: Apart from what have you said, what you would like to change to the report to make it more useful or easy to understand?

DO3: nothing. It's fine

Int: And in this case, apart from the pdf with all the issues, would you prefer one tab file with all the issues and an attribute explaining the kind of issue or do you prefer two separate tab files.

DO3: I would say one.

Appendix 21: Report feedback interview. Data owner n°5

REPORT FEEDBACK FROM DO5
July 27

Questions about basic report:

Int: In general, what do you think about a report like this:

DO5: It would be very useful. I referred it to my colleagues and they told me they had a particular project to solve this issue, but it was done identifying the problems manually.

That´s less useful [5th example from left to right and top to bottom], to me that doesn't look really as a mistake, and this one [4th example from left to right and top to bottom] is very minor. Three of the six appear to be an issue.

Int: Do anyone on your team do some kind of spatial query for analysis purposes? For example, generating a list of all the buildings that are inside a conservation area?

DO5: If we knew how to do it, that would be very useful. Because, if you own your home, there are certain things that you may want to do in your home that need to be permitted. For example, to change the windows, or fences. We would need to send a letter to all the buildings that are part of a conservation area to avoid them to make changes without permission.

Int: In that case, it's important to be aware that if you query "which are the buildings that are within each conservation area?" these buildings [4th and 5th example from left to right and top to bottom] is not going to appear in the list, as for the computer "within" is only something that its completely inside, and these buildings have a very tiny part of them outside.

DO5: Oh, so then it's important to correct them. And it would be very time saving to do this kind of things.

Int: In the 2nd page, the examples are shown using different styles. What do you think about the following aspects of the maps:

Presence of base map?

DO5: Yes, definitely. It helps to recognise the location.

Int: Presence of cartographic text in the base map

DO5: I think is helpful to find where it is.

Int: And the bottom examples have the conservation area in the Earthlight colour while the upper ones have them in a translucent grey, to represent them as context layers. Which one do you prefer?

DO5: the colour currently in use in Earthlight is better because I'm used to it. It should be translucent, anyway, to see the basemap

Int: And about the colour and size of the features being checked?

DO5: the red is saying that there is a problem? I think that´s good. But the labels are not helpful.

Int: and is the data inside the tables the one that you would want to see there?

DO5: I don't need the TOID, I need the address or postcode. The name of the CA is useful but not the CA_ID, we don't use it anymore.

Int: What do you think about the alternatives suggested to solve the issue?

DO5: I think it's helpful

Int: Apart from the report, how you would like to receive the data of the features with issues?

DO5: Probably as a layer on Earthlight because there is where I´m going to use it.

Questions about detailed report: (Give them a time to check the report alone, without explanation)

Int: Do you think that a more detailed report like this improves the usefulness of the report as a tool to improve the quality of your data?

DO5: It's funny but not really. The concept of severity doesn't work on us because in this case it should be absolute, how much is in or out doesn't matter. We are not interested in what percentage is inside. That one [2nd medium severity example], it's in or out? If it is both, that's a mistake. The severity doesn´t help in this case.

Int: And what happens in the cases flagged as low severity?

DO5: That might worth to show it because that's clearly in or out, but if we want to do a list of buildings that are in that would be a problem as we discuss before. But that would be a lower priority.

Int: And to define the conservation area boundary, what do you need to snap to? To buildings only or other things?

DO5: A conservation area should snap to a boundary line. They either goes in the middle of the road. If it bounds a housing, you go following the fence. So, it normally follows the land registry boundaries and that normally can be found on the ground as fences or walls.

Int: How you would call this type of issues? How word you would use instead of severity?

DO5: I think they are two different type of issues, first, when the building is cut in half, or in pieces and second, when the boundary line doesn't seem to be following the land registry boundary. So, I would put together the high and medium severity as the first type of error, and the low severity just as another type of issue.

And for the first type, for example in this one [2nd example of high severity] it's possible that the building is newer than the boundary, so I need to check it. If the building is newer than the boundary the boundary should be updated.

Int: Does this change the way you can use the data and correct these issues?

DO5: Yes, because for the first type of issue further research is needed to have a clearer understanding where the boundary ought to be and if the problem is because the building is newer than the boundary or due to an error when adding the data. But for the second one, I know that is a problem on how the boundary was defined.

Int: And how you would correct that?

DO5: Probably manually I would change them.

Int: So, then the chart should show only two bars, one with the buildings cut by the conservation area boundary and the other one with the buildings that show that the conservation area is not following it boundaries.

DO5: right

Int: And what do you think about the graphic way to explain the criteria to define the types of issues? is clear and easy to understand? We should change it to only two categories, of course with a different name as well, and combining the two central areas.

DO5: I don't really understand that to be honest… [he stayed looking at it a while]. Does this [the section more to the left] means that a 5% of the building is inside the conservation area while this [the section more to the right] that a 95% is inside and the other small part is outside?

Int: exactly.

DO5: I think it is excessive data, not very useful, don´t waste time on it because the examples help to understand them much better and easier.

Int: Apart from the report, how you would like to receive the data of the features with issues? As one dataset, or earthlight layer as you preferred, or one for each type of issue?

DO5: I think two layers, because they are different errors

And looking again, the chart is nice, it gives you an idea of the scale of the problem and help you to prioritize. The chart would be useful as well in the basic report.

Int: In the basic report I didn't put it because the total number of buildings was too big that the number of issues would be insignificant

DO5: but anyway, I think is good to have it.

And it's interesting. Seeing the general map more in depth, there are some patterns, some boundaries with a lot of issues and others with very few. Maybe it is related with the date of the conservation area or with the person who did it. Maybe it would be nice to have the % of issues per conservation area.