Using semantic annotation in building databases to improve information and energy modelling: a use-case of UK domestic time-series data.
working paper

Gokhan Mevlevioglu[a,c], Sukumar Natarajan[a,c], Julian Padget[b,c]

[a] *Department of Architecture and Civil Engineering*
[b] *Department of Computer Science*
[c] *University of Bath, Bath, BA2 7AY, UK*

Abstract

There is a increasing interest in modelling stock-level (i.e. local authority, regional or national) energy flows in buildings (both domestic and non-domestic) primarily as a means for technological and economic assessment of carbon abatement options. Modelling stock level building energy flows is a complex endeavour that requires the bringing together of a range of different data-sets (climate data, physical building data, occupant profiles, system profiles etc.) each with its own particular data-structure. Typically, this process can be time-consuming, repetitive and difficult to update. As new data is continually being produced, models can quickly become out-dated. We propose a semantically annotated database via an over-arching ontology that radically simplifies this process providing powerful new techniques to combine data-sets and query them. We demonstrate this technique through building up a full time-series of English Housing Survey (EHS) data (from 1970 onwards) which are not directly compatible due to changes in survey methodologies over time. We then use the combined data-set to build up a picture of changing SAP levels for new buildings over this period and plot them against mandated changes to the building regulations. The key demonstration here is the speed and the efficiency of the

Email addresses: gokhan.mevlevioglu@bath.ac.uk (Gokhan Mevlevioglu), s.natarajan@bath.ac.uk (Sukumar Natarajan), j.a.padget@bath.ac.uk (Julian Padget)

process rather than the data itself.

Keywords: residential energy, stock models, semantic, database, EHCS, ontology

---

1. Introduction

Since the development of the semantic network model in late 1960s by [1] and Collins et al (1969), data semantics and ontologies have been used in various domains such as medicine, energy and social networks. These usages have provided efficient frameworks for easy access and accurate searching tools to the relevant materials for researchers and end-users. More recently, researchers undertaking energy modelling in a range of disciplines have begun to use semantics. For example, Keirstead et al.(2009) developed SinCity (Synthetic City) energy modelling systems at Imperial College in order to use urban energy systems (UES) ontology in a context involving agent based modelling (ABM). Similar ontology to UES, which is called socio-technical systems (STS), was created by Van Dam (2009) for Delft University of Techology to help development of ABM. Later,Van Dam & Keirstead (2010) researched usage of UES and STS ontologies for urban energy system modelling. They analysed ways of ontology translations in order to demonstrate interoperability of ontology based energy systems. However, a similar technique has to-date not been applied to domestic housing databases. The primary purpose of this paper is to reveal the benefits of utilising ontology based semantics to building energy modelling through the use case of the UK domestic housing stock.

In order to demonstrate the benefits of semantic UK housing database by addressing these issues, following steps were taken and explained:

- General information about UK housing databases, namely English House Condition Survey (EHCS) and English Housing Survey (EHS) was given along with details of existing studies in background section.
- Creation of sample semantic UK housing database from current major relational datasets from year 1986 to year 2008 was explained in method-

ology section.

- Development of data ontology and shared semantic language to structure the semantic database was outlined in methodology section, so that database can be queried efficiently regardless of the source and year of database (i.e. EHCS-1986 or EHS-2007).
- Development of an application to query the semantic database was explained in methodology. The application also presents the query results in graphs for comparisons between different years and surveys.
- Carrying out research on insulation improvements of the UK housing in sample regions with semantic database and querying application. This research was illustrated in case study 1.
- Case study 2 demonstrates the usege of this research method to include datasets from different domains, such as gross domestic product information datasets or energy policy implementations. This study also supports interoperability of datasets that are based on ontologies.

Creating database and using the ontology to give extra dimension to the database would not be useful for any research without a tool to query this newly created complex database. Because this database type is not common for housing data, there were no ready applications that can query this database efficiently. Consequently, a new Java application was created in this study to use with the database. The application, which is named "SemUK", has features such as querying longitudinally, querying transversely, simultaneous querying, creating result tables and producing charts of results. Figure 5 displays the work-flow of the application briefly as well as generic features of it.

SemUK was then used in UK housing research related with insulation measures and quality, to demonstrate the speed, efficiency and query accuracy of the system. It was also possible to add datasets of different domains like GDP and Energy Efficiency Commitment (EEC)-Carbon Emissions Reduction Target (CERT) policies. Adding different data domains allowed comparisons among the query results to identify relationships between different data events, which would take tremendous amount of effort with a research approach based on

solely relational datasets. Finally, other potential usage implications such as integration with a universal ontology to create stronger tool or using database with other tools like Domestic Energy and Carbon Model for the UK (DECarb) were identified.

2. Background

Survey files used for this study, UK housing stock surveys EHCS and EHS, can be accessed through Economic and Social Data Service (ESDS). They are stored separately for time series and for different relational data analysers such as IBMs SPSS software or Statacorps STATA software. While it is possible to analyse either EHS or EHCS data in different time series for specific variables with these software at the cost of great effort and substantial time, it is not possible to analyse them together when variable names and values change over the time. For example, one survey may classify construction variable value as "masonry", while another survey may decide that they should classify "single leaf masonry" and "cavity walls" separately. Or, their naming standard system for variables can have variation. For instance, construction variable in a survey can be called "constr" and it can be "constrmatl" in another. Eventually, small differences like these make these closely related datasets incompatible, resulting in the lost possibility to analyse them together with statistical software built for relational databases.

For the same reason of incompatibility, it is not possible to extend the scope of these databases by adding different surveys administered by other institutions. That is to say, some surveys may have specialised subject that includes variables that other surveys do not have. If the databases of these surveys were compatible, it would be possible to broaden these databases with different variables in other databases. Furthermore, housing database may also have relations with other domains such as socio-economic or climatic data domains, which have very little similarity in terms of data variables. To conduct a research analysing the relations between different domains is currently very difficult and

time consuming task as data structure of relational databases are not fully prepared to allow data integrations between different domains. Although there are currently no examples of semantic UK housing database or ontology in the literature, there are various examples in other domains.

In order to display the benefits of semantic database more clearly, problems and limitations of current relational housing datasets can be summarised as:

- Current datasets are part of different surveys, such as EHCS and EHS, which took place until today. These datasets were recorded quinquennially from 1971 to 2001 and annually from 2003 to 2009 as results of these surveys. Data classification and data identifications in these surveys differ from each other, which makes it difficult to find same type of information in different surveys.(SEE FIGURE 4)
- Because the data is stored separately for each year that particular survey was taken, it is impractical to make a chronological search that includes several years data.
- The effects of relevant domains such as household information, energy modelling and legislations cannot be analysed altogether due to a lack of a shared framework that can simulate inter-relations between these domains.

One of the closest ontology examples in literature to the one built by this study may be the one in Van Dam and Keirstead (2010)s research. Their research was related with energy systems ontologies and about how to "translate" different ontologies in an attempt to use them in an interoperable manner. Because energy systems domain had already had many different ontology applications by the time of the study, their purpose was to link those ontologies to test the feasibility of interoperability of these ontologies. They analysed two major methods of ontology translation. One of them required a master ontology, which is flexible and adaptable to existing ontologies to unite any related database, while other method was to directly translate two ontologies in question in their research. Second approach would have been quicker, as it would not require bigger perspective to take into account besides two existing ontologies. However,

they eventually had chosen to use a master ontology for the interest of community that might take this master ontology as a step through a standard ontology for energy systems modelling. The experiment in the study was successful on merging ontologies. The advantages of the approach are summarised as, firstly, possibility of re-using data from other ontologies without having to re-create them and secondly, addressing the strong need of including different kinds of data in energy systems, such as environmental, socio- economic and technical data. One set-back of the approach was explained as the need of manual data fine- tuning process, as the current number of energy ontologies creates high complexity for any attempt to standardise the energy system ontologies.

3. Methodology

Semantic annotation of data is a non-destructive method of applying hierarchical structure to a dataset by connecting the data substrate within a broad ontology. Ontology in this context is a tree of data structure, which annotates data properties according to relevant branches of the tree. This is achieved by adding "metadata" to existing data. Metadata (which can be understood as "data about data"), holds the information about the data categories and values within the database (Hay 2006).

For example, in a conventional dataset, a house sample may be recorded as being in "Inner London" where in another dataset the very same sample may be recorded as being in "London" or in "East" region. When a meta-data is implemented, which holds the information that "Inner London" is in "London" and "London" is in "East" region, an interpretation with a reasoning method can be done to negotiate between these datasets (Figure 1). The resultant negotiated database along with conventional datasets is referred as "Semantic Database."

Figure 2 shows the original context of the surveyed data in a relational database, which is then converted to a graph database with "Mapping Master" as it is shown in figure 3. As soon as mapping is completed, "A home A1002050"

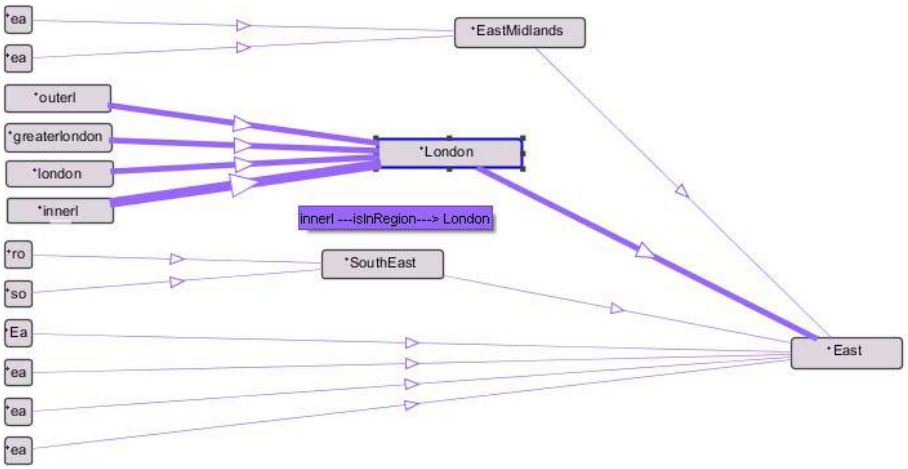


Figure 1: Example data relation in a database with meta-data

EHCS2003

	A	B	C	D	E	F	G	H	I
35	A1002044	northeast	semidetached	masonrycavity	37	gasfiredsystem	otherurbancen...	150	cavitywithinsu...
36	A1002045	northeast	purposebuilt	masonrycavity	62	notidentified-c...	otherurbancen...	121	cavitywithinsu...
37	A1002047	northeast	midterrace	masonrycavity	78	gasfiredsystem	suburbanresid...	150	cavitywithinsu...
38	A1002050	northeast	midterrace	9inchsolid	76	gasfiredsystem	otherurbancen...	92	other
39	A1002051	northeast	midterrace	masonrycavity	46	gasfiredsystem	otherurbancen...	100	cavityuninsula...
40	A1002054	northeast	endterrace	masonrycavity	70	electricalsystem	otherurbancen...	100	cavityuninsula...
41	A1002055	northeast	midterrace	masonrycavity	73	gasfiredsystem	otherurbancen...	100	cavityuninsula...

Workbook File

Figure 2: Relational database representation of “A home A1002050”

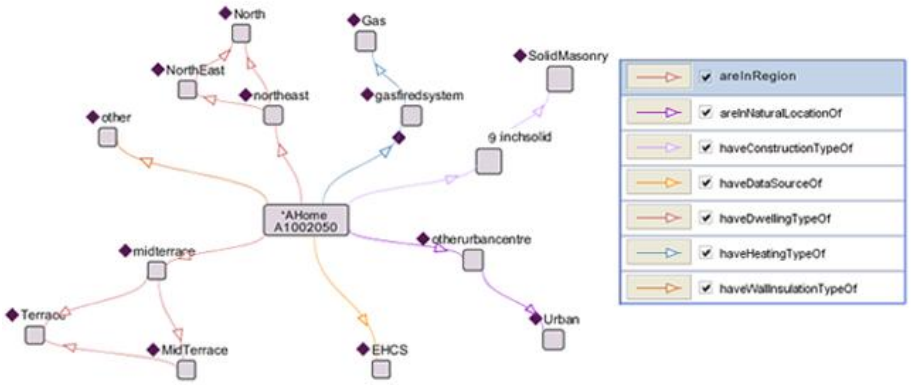


Figure 3: Graph database representation of "A home A1002050"

becomes part of a semantic database, which can be used to infer new relations based on the data mappings and ontology created. The inference can be done by process called "reasoning." For example, by using reasoning, database knows that "A home A1002050" which is located in north-east region is located in a North region at the same time. Therefore, it will be retrieved by queries asking for houses in either North region or more specifically north-east region. This property of semantic database exceeds the capabilities of relational database in a basic sense.

After creating semantic datasets by structuring the data, the challenge was to find a shared language to represent all datasets, so that they can be used together. That is where ontology approach helps by mapping different structures of data with a shared language. Figure 4 demonstrates how differently named data values in Year1986Dataset's dwelling type variable values (top window) and Year2004Dataset's variable values (bottom window) are mapped to "Shared value list" (middle window) with Protégé. Variable values with same names in different datasets exist as well in some cases (e.g. "midterrace" value). Another Protégé plug-in "Jambalaya" is used to create the visual demonstration of mappings performed. This approach was applied for all variable groups of datasets.

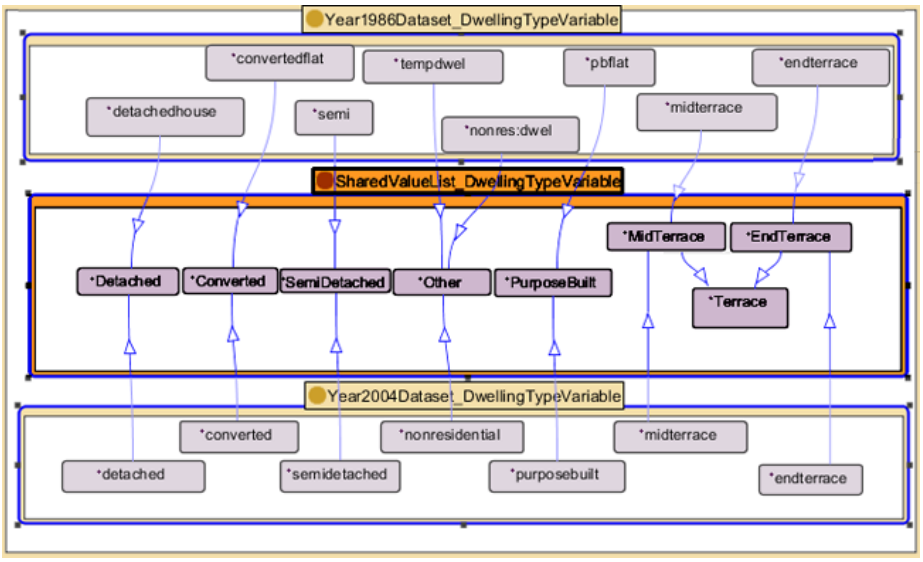


Figure 4: Example mapping of dataset values to "shared value list

Various domestic housing and energy surveys were reviewed such as Scottish Household Survey (SHS), Homes Energy Efficiency Database (HEED) and Survey for English Housing (SEH) on desk based study. EHCS and EHS datasets from year 1986 to 2009 are selected to be included in this research. Although EHCS data was available from 1971, there was not any record of data between 1971 and 1986, which does not make inclusion of 1971 data advantageous. EHCS and EHS surveys are reviewed predominantly through their associated literature followed by an explanation of survey components included in surveys.

Transforming existing housing datasets into semantic datasets required following basic steps to be carried out:

- Selection of variables to be used in semantic database ,
- Dealing with missing data and data to be drawn from existing variables,
- Importing relational database files into Protégé as graph data,
- Building the ontology based on variable values,
- Exporting database as RDF-XML based OWL file and
- Using AllegroGraph to convert database into indexed triples for efficient

querying.

The purpose of the variable selection stage was to identify key variables that can help demonstrating qualities of the semantic querying approach in a research related with energy efficiency in the UK housing. For this purpose, selected variables were related with location of the sample housing, dominant construction variable, energy efficiency related variables, dwelling type and housing unit area variables. The missing data in the survey was treated as they were and included in the database with "unknown" tag. The selected variables were then extracted from the source files and imported into the Protégé software, which is an ontology building tool. The ontology was built and implemented to the database with the help of "Mapping Master" plug in of Protégé. Then, the ontology integrated data was exported as OWL file format. This format could be read by AllegroGraph software, which would create the semantic database from data triples. Adding database from different data domains was done by the same procedure. AllegroGraph software also has capability of receiving queries and returning the results of the queries from the semantic database to the querying application.

Preparation of querying application was done using Java programming language, mainly because its capability of running in different platforms on a virtual machine. This allows the application to be accessible by more people. Availability of java libraries for AllegroGragh application user interface allowed to create seamless integration of the application with AllegroGraph. Application flow, user interaction, and AllegroGraph communications are demonstrated in Figure 5.

Application utilises lists and combo boxes to show available variables and variable values for querying in "shared value lists." Values that are not in the list can be included in the query through typing in combo box fields if value tags are known by the user. Tables are used to display and store query results. Charts can be added to tabbed pane with chart buttons for analysis (Figure 6). The built application was ready to be used in research related with the UK housing stock.

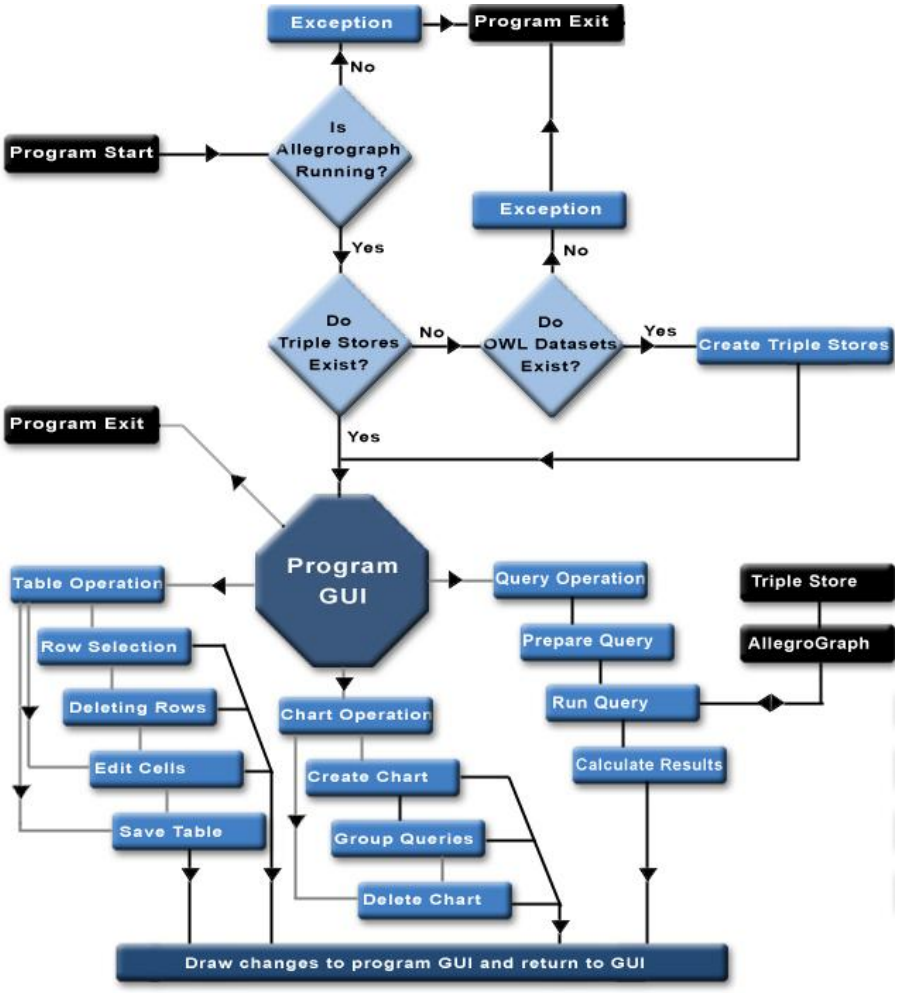


Figure 5: Application flow-chart

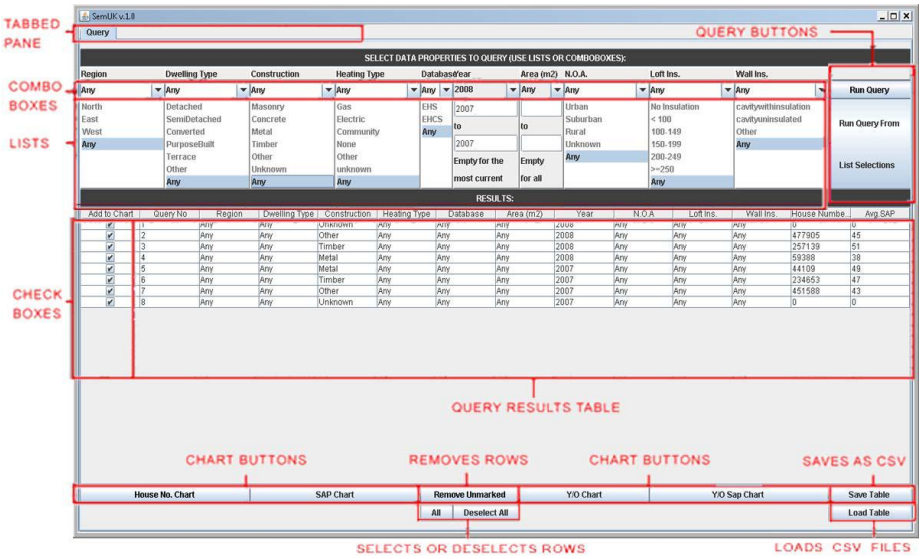


Figure 6: Graphical user interface of the application showing sreen components

Finally, accuracy, efficiency and versatility of the prepared application was tested through various queries regarding the insulation properties of UK housing research.

4. Case Studies

Two case studies were presented in this paper to see the functionality of the approach. The purpose of the first study was to demonstrate the analysis potential and functionality of the SemUK application in a research regarding single domain. The second study on the other hand, was aimed to address the research capabilities of the application in more than one domain. On the other hand, it was also important to select an area that represent substantial portion of the UK housing stock.

4.1. Case Study 1: Longitudinal Research on loft insulation improvements

The research area for this study was selected to illustrate the complex querying functions of the application. A precedent analysis is undertaken to define

Query No	Type	Area (m2)	Year	N.O.A	House No
1	**Detached**	Any	2008	Any	4,992,676
2	**Terrace**	Any	2008	Any	6,628,733
3	**SemiDet.**	Any	2008	Any	6475395
4	**Converted**	Any	2008	Any	792323
5	**P.Built**	Any	2008	Any	3317361
6	**Other**	Any	2008	Any	32910
7	Any	**< 50**	2008	Any	2,480,232
8	Any	**50-69**	2008	Any	5,346,676
9	Any	**70-89**	2008	Any	6,378,625
10	Any	**90-109**	2008	Any	3,187,750
11	Any	**>110**	2008	Any	4,846,115
12	Any	Any	2008	**Urban**	4,931,130
13	Any	Any	2008	**Suburban**	13,191,686
14	Any	Any	2008	**Rural**	4,116,582

Figure 7:

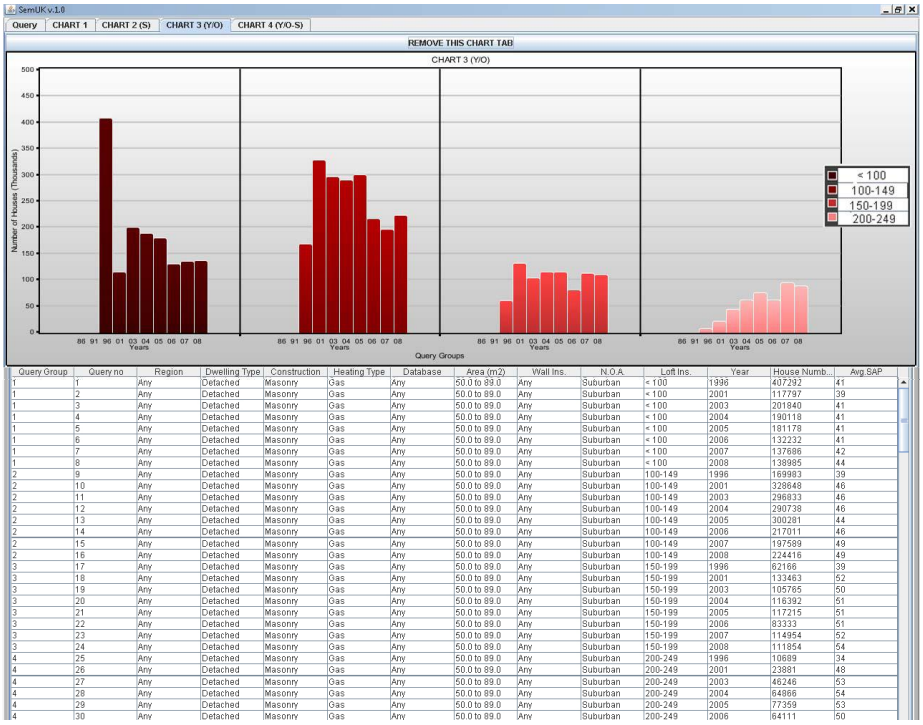
house properties, which represents large part of the stock in order to select which properties to be used. Properties that represent large number of houses are selected because they would be better to check the accuracy of the query results as more housing samples are involved.

Table 1. Precedent query results by the SemUK application

Masonry wall structure and gas heating types are already known to be dominant part of the housing stock based on the EHCS 2009 report. Therefore, they are not included in the precedent queries. First six queries of Table - 1 showed that "Detached" and "Terraced" dwelling types are prominent types. Although terraced houses are more, detached houses are selected for the research. This is because terraced houses included two other types; mid-terrace and end-terrace types in that number. Next four queries (7-11) showed that houses, which have area in between 50-89 m^2 have nearly as half as many as the housing stock. Final three queries showed that most of the housing stock is located in suburban areas. Considering these points, it seemed to be reasonable to select houses with properties in Table - 2 for the research.

Region	Dwelling Type	Construction	Heating Type	Database	Area (m_2)	Year	N.O.A	Loft Ins.	Wall Ins.
Any	Detached	Masonry	Gas	Any	50.0 to 89.0	1996 to 2008	Suburban	Each Band	Any

Figure 8:



Query Group	Query no	Region	Dwelling Type	Construction	Heating Type	Database	Area (m2)	Wall Ins.	N.O.A	Loft Ins.	Year	House Numb...	Avg SAP
1	1	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	< 100	1996	407292	41
1	2	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	< 100	2001	117797	39
1	3	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	< 100	2003	201840	41
1	4	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	< 100	2004	190118	41
1	5	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	< 100	2005	181178	41
1	6	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	< 100	2006	132232	41
1	7	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	< 100	2007	137686	42
1	8	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	< 100	2008	138985	44
2	9	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	100-149	1996	189983	39
2	10	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	100-149	2001	328648	46
2	11	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	100-149	2003	296833	46
2	12	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	100-149	2004	290738	46
2	13	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	100-149	2005	300281	44
2	14	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	100-149	2006	217011	46
2	15	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	100-149	2007	197589	49
2	16	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	100-149	2008	224416	49
3	17	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	150-199	1996	62186	39
3	18	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	150-199	2001	133463	52
3	19	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	150-199	2003	105765	50
3	20	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	150-199	2004	116392	51
3	21	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	150-199	2005	117215	51
3	22	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	150-199	2006	83333	51
3	23	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	150-199	2007	114954	52
3	24	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	150-199	2008	111854	54
4	25	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	200-249	1996	10689	34
4	26	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	200-249	2001	23881	48
4	27	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	200-249	2003	46246	53
4	28	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	200-249	2004	64866	54
4	29	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	200-249	2005	77359	53
4	30	Any	Detached	Masonry	Gas	Any	50.0 to 89.0	Any	Suburban	200-249	2006	64111	50

Figure 9: House numbers chart of initial analysis for loft insulation bands (from 0 to 250 mm)

Table 2. Initial selection of house properties for research

Since the study is related with the energy efficiency properties of the UK housing, the loft insulation properties are selected to be broken down for houses with “detached” dwelling type, “masonry” construction type and “suburban” nature of location with the use of SemUK application.

After querying the houses with the selected variables and plotting the chart from the query results practically with SemUK, a curious point stood out in the analysis. There is an unexpected decrease in every loft insulation band in year 2006 survey. In order to find the reason of these decreases, more queries

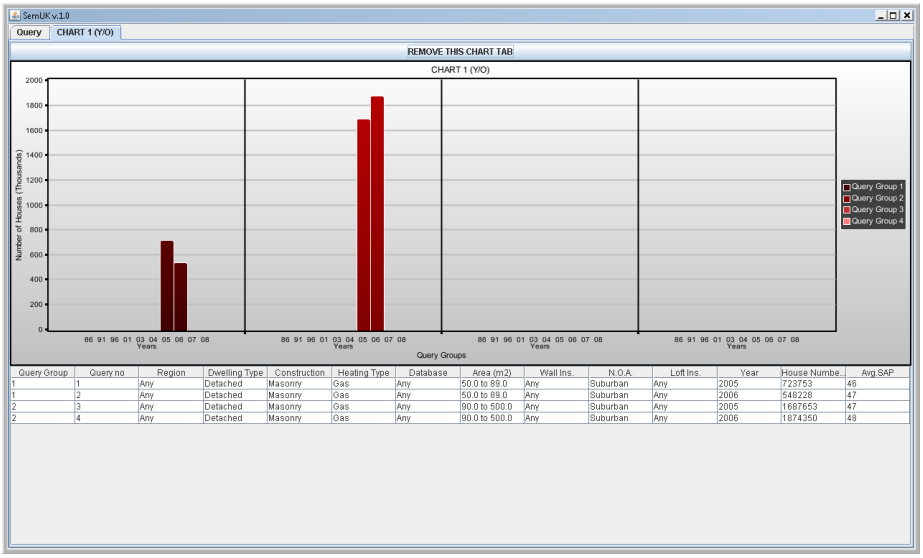


Figure 10: House numbers of 50-89 m2 (Query Group 1) and 90-500 m2 (Query Group 2) categories

are undertaken and it is found that the source property of the decline is area of houses. House number of houses with 50 to 89 m^2 floor area decreased while bigger houses number increased in 2006 (Figure 7 and Figure 8).

(Requires source information)

It was later confirmed that the annual report of EHCS 2006 issued a chart showing changes in housing stock, which involves increasing floor areas. Figure 9 shows that significant percentage of houses increased their floor area in that year and consequently resulted a decrease in total dwelling numbers of the queried types. This unexpected change in houses sizes could interfere with the interpretation of the analysis results. Since the research is aimed to find out the influence of loft insulation thickness to certain types of houses, the resultant influence in the research might be caused by the changes in house areas. In order to minimise these unexpected interference to the results caused by changes across the property categories in time series, "area" and "heating type" properties were not included in the research.

After a precedent analysis with the application, housing properties that

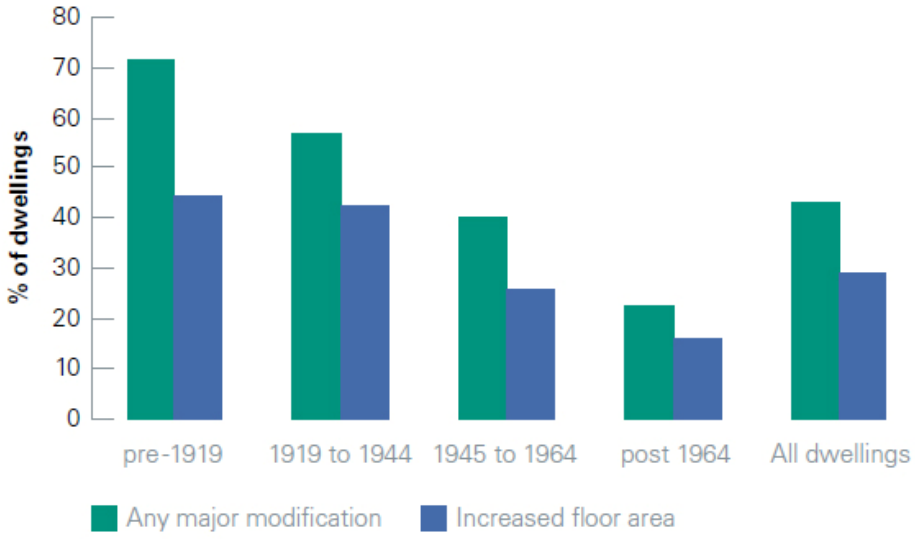
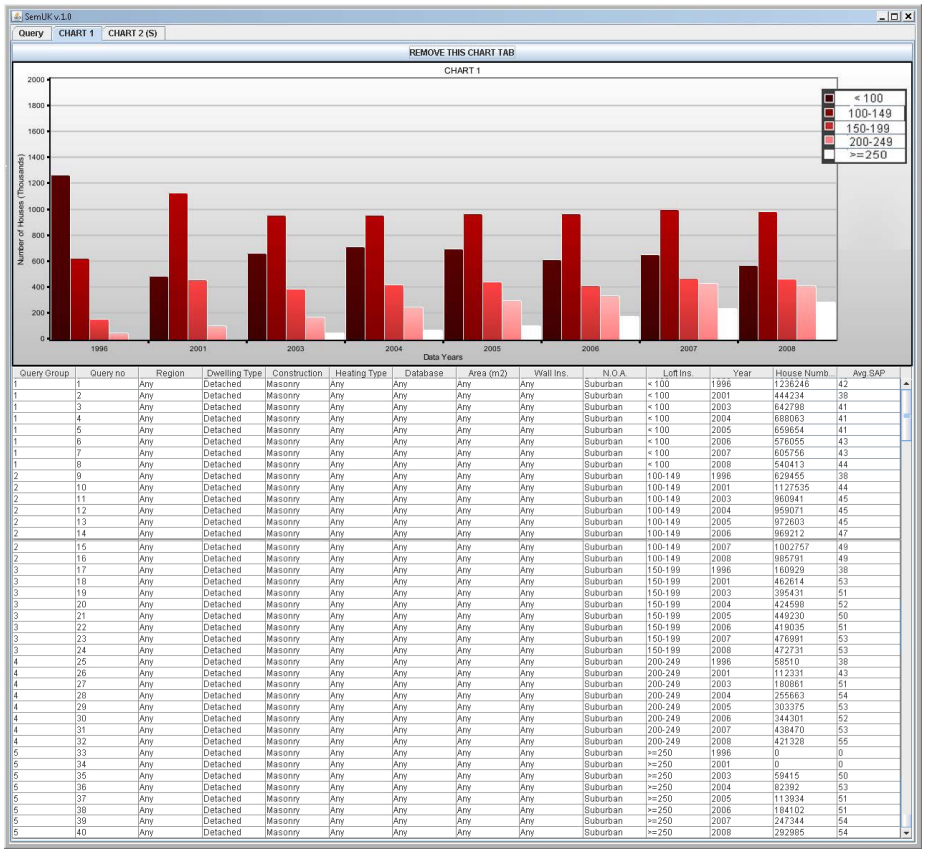


Figure 11: Percentage of dwellings with any major modifications and work to increase floor area

would encompass higher number of houses without data interference were identified again as "masonry" type housing construction, "detached" type dwelling and "suburban" type nature of location. By being able to query and analyse the selected properties fast enough, it was possible to change the selected variables for the research in the early stage.

The queries were set up with these properties for the range of survey years and each loft insulation band, results are gathered from application results table of the graphical user interface. Chart tabs were created from that table for different type of charts.

It can be seen from Figure 10 that, there is a similar pattern of insulation bands for each survey year from 2001 to 2008. According to this pattern, most houses in detached, masonry construction and in suburban areas have loft insulation thickness between 100 and 149 mm while second most common thickness band is in between 0 and 99 mm. "150-199", "200-249" and "250 and more" thickness bands are following them respectively. Only exception to this pattern is in year 1996 survey data, where 0-99 mm band has the most number



Query Group	Query no	Region	Dwelling Type	Construction	Heating Type	Database	Area (m2)	Wall Ins.	N.O.A.	Loft Ins.	Year	House Numb.	Avg.SAP
1	1	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	< 100	1996	1236246	42
1	2	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	< 100	2001	444234	38
1	3	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	< 100	2003	642798	41
1	4	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	< 100	2004	688063	41
1	5	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	< 100	2005	659654	41
1	6	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	< 100	2006	576055	43
1	7	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	< 100	2007	605756	43
1	8	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	< 100	2008	540413	44
2	9	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	100-149	1996	629455	38
2	10	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	100-149	2001	1127535	44
2	11	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	100-149	2003	960941	45
2	12	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	100-149	2004	959071	45
2	13	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	100-149	2005	972603	45
2	14	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	100-149	2006	969212	47
2	15	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	100-149	2007	1002757	49
2	16	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	100-149	2008	985791	49
3	17	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	150-199	1996	160929	38
3	18	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	150-199	2001	462614	53
3	19	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	150-199	2003	395431	51
3	20	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	150-199	2004	424598	52
3	21	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	150-199	2005	449230	50
3	22	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	150-199	2006	419035	51
3	23	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	150-199	2007	476991	53
3	24	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	150-199	2008	472731	53
4	25	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	200-249	1996	58510	38
4	26	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	200-249	2001	112331	43
4	27	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	200-249	2003	180861	51
4	28	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	200-249	2004	255663	54
4	29	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	200-249	2005	303375	53
4	30	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	200-249	2006	344301	52
4	31	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	200-249	2007	438470	53
4	32	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	200-249	2008	421328	55
5	33	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	>=250	1996	0	0
5	34	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	>=250	2001	0	0
5	35	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	>=250	2003	59415	50
5	36	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	>=250	2004	82392	53
5	37	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	>=250	2005	113934	51
5	38	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	>=250	2006	184102	51
5	39	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	>=250	2007	247344	54
5	40	Any	Detached	Masonry	Any	Any	Any	Any	Suburban	>=250	2008	292985	54

Figure 12: House number chart of the research for loft insulation levels from “¡100” to “¿=250” mm

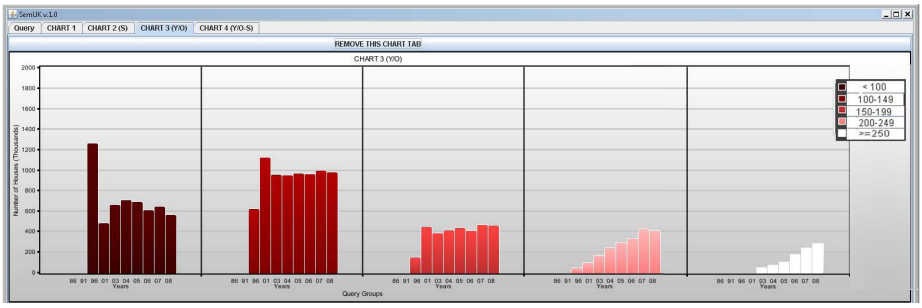


Figure 13: Year ordered chart for loft insulation levels from “¡100” to “¿=250” mm

of houses. It is easier to see the changes of the numbers for each band across the years in “year ordered” house number chart. Year ordered chart in Figure 11 and succeeding SAP charts do not have query table underneath, as they are same as house number charts query table.

In Figure 11, two distinctive patterns can be observed. First pattern has fluctuating start and relatively steady end, which can be observed in “¡100”, “100-149” and “150-199” mm band graphs. The amount of sudden decrease in “¡100” category is almost same as the total of sudden increase in “100-149” and “150- 199” mm categories in 2001. This suggests that houses in “¡100” category mainly changed their insulation thickness to other two categories in 2001. Second pattern is the steady increasing pattern of 200-249 and ¿=250 mm bands. Figure 34 shows that there is also a steady increase in total house numbers for aggregate of all loft insulation bands. It can be inferred that the majority of newly built detached houses with masonry construction in suburban areas had loft insulation band of 200 mm and above.

Average SAP ratings of the houses in research showed different patterns for each loft insulation band throughout the survey years analysed (Figure 12). It can be noticed that change trend is towards the improvement of the SAP ratings. The most consistent improvement is in “100-149” mm insulation band. Average values for this band never displayed decrease, while there are upward and downward changes in other insulation bands. The biggest improvement is in “200-249” mm band. Average rating increased to 55 in 2008 survey from

Figure 14: SAP rating chart for loft insulation levels from "¡100" to "¿=250" mm

38 in 1996 survey. On the other hand, the least improvement is in "¡100" mm group with 3 points change during the survey period analysed in research.

This case study demonstrated that successful creation of semantic database with AllegroGraph software proved the feasibility of applying semantics to the UK housing domain. Fast querying performance of SemUK provided house properties that will interfere with the results to be detected quickly and because of that it was possible to shape research extents accordingly. It was easy to focus the queries to the year 2005 and 2006 datasets, where unexpected declines in house numbers occurred. The easy accessibility attribute of semantics rendered fast focusing possible with disambiguated querying attribute.

It was possible to see the effects of the ontology and shared language after the development of a semantic query application, which was written in Java language. The application was accurate in terms of retrieving the queried results from different surveys database for consecutive years. Besides, the results were obtained through same queries for different data in a time efficient manner. Queries did not have to include all the different terminology used in different datasets for same materials, as shared language and ontology implementation was interpreting the query for different terminologies. The research done with the application demonstrated that newly built homes had higher insulation thickness, which are mostly higher than 200 mm in loft areas. It was also clear from the research that there is a strong relation between insulation and SAP ratings, while it suggested that this relation was not as strong for the homes

that have thicker insulation levels as the homes with thin or no insulations.

4.2. Case Study 2: Transversal integration of domain data

The purpose of this case study was to demonstrate the data integration from different domains to the semantic database. The expected outcome of this type of integration was to see the relations of events in different data domains that affects the main housing database we have created. For example, a sudden decrease in population's income level in a region might have had influence on how they upgrade the building envelope of their home to benefit from fuel expenditure reductions. Following data domains were added to the semantic database to demonstrate this effects:

- GDP data of nine governmental regions of England between 2001-2009 (Eurostat, 2012)
- CERT-EEC loft and cavity wall insulation numbers data from 2003-2012 (EST-HEED, 2012)
- Fuel Poverty data from 2003 to 2009 (DECC, 2012)
- Domestic Fuel Consumption data from 2003 to 2009 (DECC-2, 2012)

All of the datasets were detailed to governmental region level, which made it easier to integrate them to current database and query them together. GDP datasets included total gross domestic product for governmental regions for the specified period. From CERT-EEC datasets, specifically loft and cavity insulation measures extracted. These measures are the number of insulation measures implemented to houses in governmental regions due to CERT-EEC policies implemented. Fuel poverty data shows the number of households, who are in fuel poverty in regions. Domestic fuel consumption data shows the total fuel consumption in regions for domestic use.

Using all datasets, queries were set to retrieve results for each region in each data category for analysis. The results were projected on charts for loft insulation and cavity wall insulation separately for each nine governmental region. Aggregated results for east, west and north regions were also produced. Initial analysis yielded interesting result for the houses in London region.

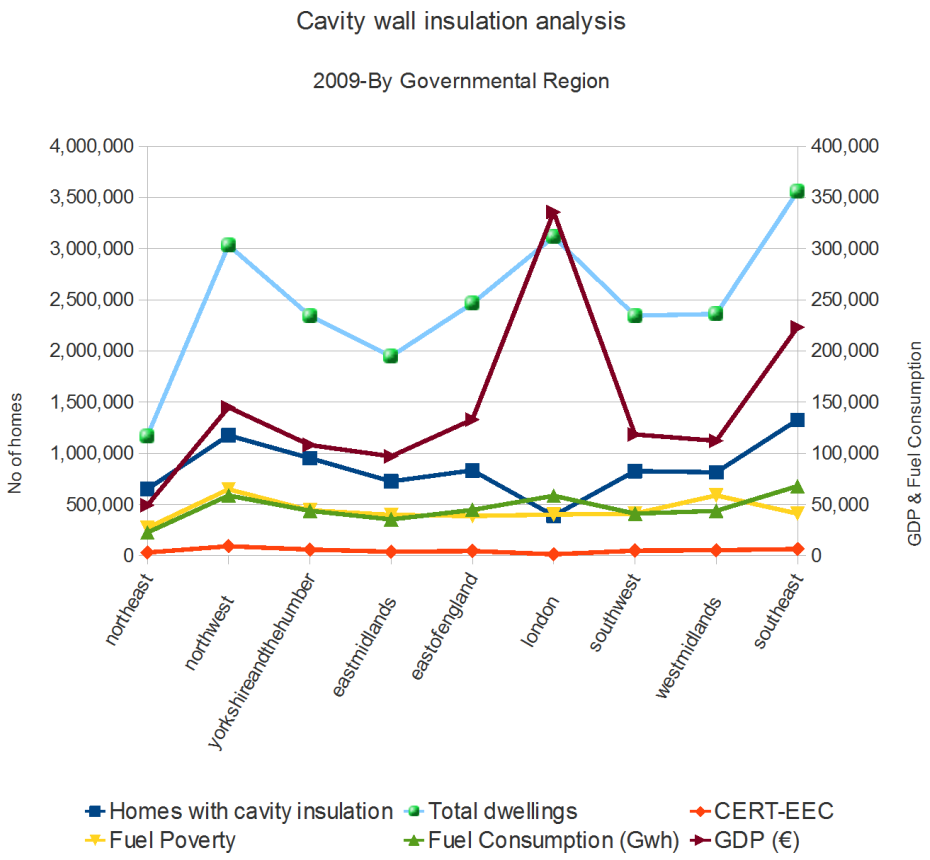


Figure 15: Cavity wall insulations by governmental region

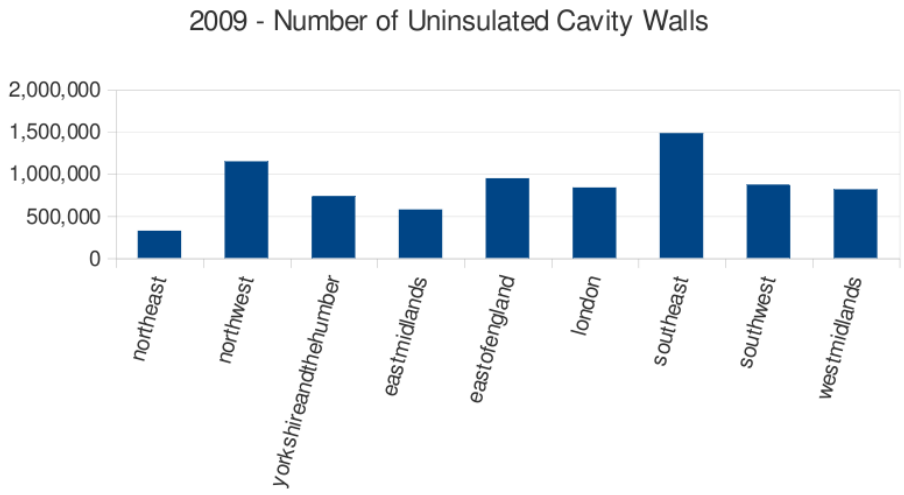


Figure 16: Uninsulated cavity walls in 2009

Figure 13 shows that total GDP in London region dominantly high compared with other regions, while total number of homes with cavity insulation are significantly less. Because climatic conditions in London is comparatively better than northern regions, one would expect that total domestic fuel consumption would be low too. However, fuel consumption in London area was virtually as high as in north west region, which has similar total number of dwellings. Also, north west region has as thrice as many dwellings with cavity wall insulation as London.

These results suggest that dwellings' energy efficiency levels in London are not very high and there can be big potential of energy savings due to fuel consumption in London dwellings. It can also be seen from the figure that the least CERT-EEC measures were implemented in London region, which suggests that CERT- EEC measures are not very effective to cause dwellings to insulate their cavity walls. With further querying, it was found that there were 837,556 dwellings with uninsulated cavity walls (Figure 14). Queries of loft insulation levels pointed out a similar result.

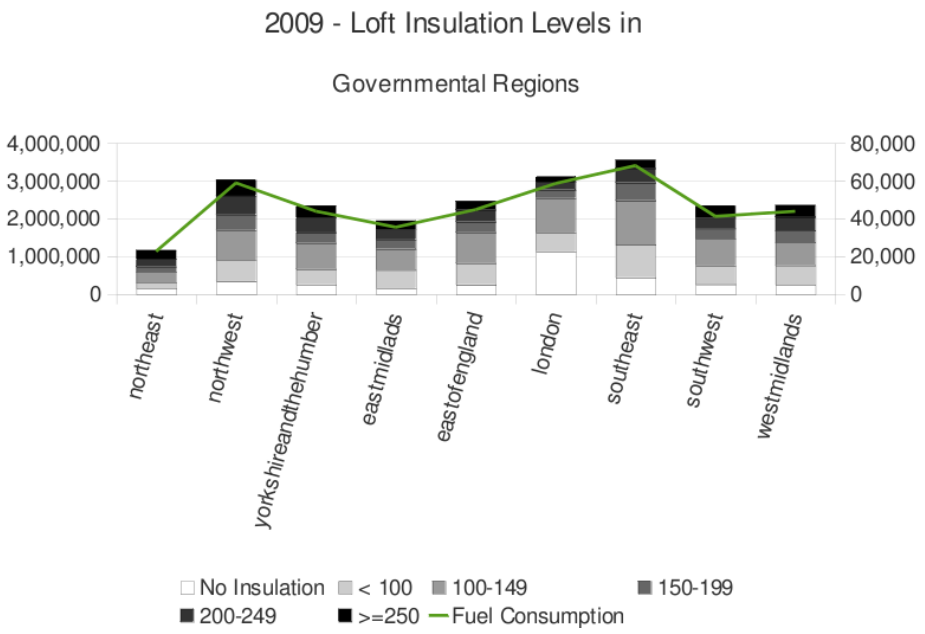


Figure 17: Loft insulation levels chart in 2009

Figure 15 shows that loft insulation levels in London region are lower than other regions as light colour stacks in chart are more than dark coloured stacks for London. This result also supports that the insulation upgrades in the region are not as many as in other regions. However, fuel consumption levels are changing proportionally with the number of dwellings regardless of level of insulation or climatic conditions of different regions. This result questions the effectiveness of insulation upgrades in terms of fuel economy. Besides the study results, there were also unexpected findings regarding the EHS and EHCS datasets.

During the analyses of cavity wall insulations' improvements in regions, unlikely results were observed in East Midlands and East of England regions. The results were showing that houses with cavity wall insulation reduced gradually after 2007 until 2009. Figure 16 shows that there were 150 thousand houses less in 2008 compared with 2007 in East of England. CLG (2012) figures show that there were about 22,000 new buildings finished and 22,000 new building constructions started in region. In order to justify the EHS data results, there

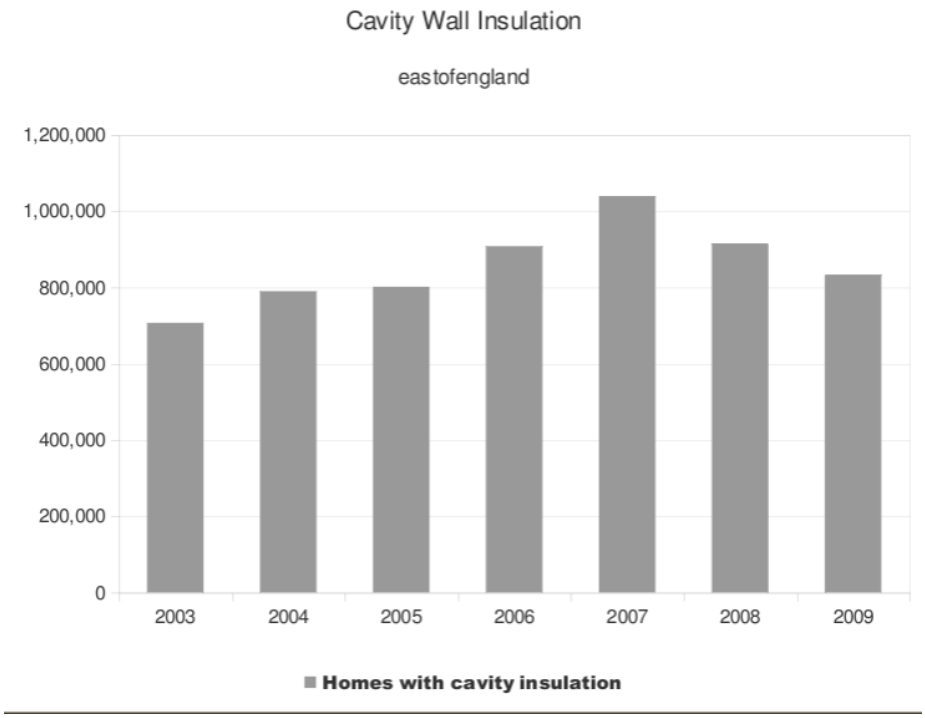


Figure 18: Homes with cavity wall insulation in East of England

should have been about 172 thousand demolitions in London in 2007. However, Boardman (2005) stated that average demolition rate in total in England was about 20,000 per year. This information leaves out one option that about 170 thousand dwellings in 2007 removed their cavity wall insulation, which is highly unlikely.

Further analyses showed that there were similar unlikelihood in other parts of the database, such as total number of houses in east regions. SemUK application with semantic database, allowed highly detailed research in time series, which gave opportunity to analyse the data very deeply. Consequently, valuable results that was "hidden" in the complexity of the collected data as well as "hidden" anomalies within the data were revealed with a relatively short study.

5. Conclusions and suggestions for future works

Sample ontology and semantic database demonstrated the potential benefits of using them such as efficient querying and ease of accessibility to data. The results suggest that an implementation of semantic database in large scale can open possibilities of better analysis and more comprehensive research related with housing in a shorter research time compared with research done with relational datasets. The benefits demonstrated in this paper can be summarised in three sections:

- Research efficiency by integration of dataset longitudinally,
- Comparability of datasets of different domains for certain time series and
- Unification of related datasets to be queried together for comprehensive analyses.

What provided the efficiency of the research was that the query results acquired from longitudinally integrated datasets (or vertically; same datasets from different years) were produced fast enough to allow more time on data analyses. Eventually, this leaded to possibility of detecting abnormal data, which is produced via weighting of survey sample. On one hand, longitudinal implementation had "catalyst" effect to improve research quality faster, on the other hand, great potential benefits were suggested by the transverse (horizontal; different dataset domains implemented for similar time series) implementation of different database domains.

Comparability of datasets were achieved by introducing different domains of data, namely fuel poverty, GDP, EEC-CERT and fuel expenditure to the system. One of the main advantages one would expect of having different datasets queried together through a shared ontology is to find out how events of different domains affect each other. By comparing different data in this paper, it was demonstrated that London region has low implementation rate of insulation upgrades, although it has significantly higher GDP levels than other regions. However, the energy expenditure reduction effect of insulation upgrades was not seen in the results, which can be a topic for a new research.

Unification of the similar datasets has two major advantages. Firstly, it allows increasing the longitudinal coverage of the datasets by adding EHCS datasets and EHS datasets to be queried together, thus extending to a larger historical data. Secondly, more detailed data can be acquired by adding more variables to the system from different datasets for same time series. For example, more housing construction variables can be added to the system from other data sources to include houses with alternative construction, such as adobe construction.

Besides the demonstrated advantages and advantages that are implied by the results of this study, there are suggestions, which can be basis of further studies. For instance, universal ontology that would integrate the housing domain with household information or energy policies can make it easier to see the effects of policies on the population and eventually on houses. The model implementation of EEC-CERT policy is a promising example that it is feasible to do so. Using a universal ontology system that may allow improving ontology itself has potential to create a huge web of data, which can be queried accurately. It also opens doors for using this system in new applications that can take advantage of any part of this huge web of data. Currently, implementation of the database can be used with other Java based applications. For example, a future housing energy efficiency tool DeCARB can use the database to predict the future condition of the housing stock. In this aspect, the expandability of the database and application was shown in the research, which allows customising the query application and expanding the database with new surveys over the time to keep the database up to date, or improve its qualities. With the help of appropriate tools of programming, it is not difficult to implement the applications to a world wide web application, which can reach more researchers and population.

6. References

Boardman, B et al (2005), 40 percent house, Environmental Change Institute, University of Oxford, Oxford.

Collins, et al (1969). "Retrieval time from semantic memory". Journal of verbal learning and verbal behavior 8 (2): 240247.

CLG (2012). House Building: December quarter 2011, England [Online]. Available from :

`http://www.communities.gov.uk/publications/corporate/statistics/housebuildingq42011` [Accessed: 21 May 2012].

DECC (2012). Fuel Poverty Statistics [Online]. Available from:

`http://www.decc.gov.uk/en/content/cms/statistics/fuelpov_stats/fuelpov_stats.aspx` [Accessed: 20 May 2012].

DECC-2 (2012). Energy consumption in the United Kingdom [Online]. Available from:

`http://www.decc.gov.uk/en/content/cms/statistics/publications/ecuk/ecuk.aspx` [Accessed: 21 May 2012].

EST- HEED (2012). Energy Saving Trust-Introduction to HEED [Online] Available from:

`http://www.energysavingtrust.org.uk/Professional-resources/Existing-Housing/Homes-Energy-Efficiency-Database/Introduction-to-HEED` [Accessed: 20 May 2012].

Eurostat (2012), GDP at regional level [online]. Available from:

`http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/GDP_at_regional_level` [Accessed: 21 May 2012].

Gruber, T.R.(1993). Toward principles for the design of ontologies used for knowledge sharing. Presented at the Padua workshop on Formal Ontology, March 1993, later published in International Journal of Human-Computer Studies, Vol. 43, Issues 4-5, November 1995, pp. 907-928

Hay, C. D. (2006). Data Model Patterns: A Metadata Map (The Morgan Kaufmann Series in Data Management Systems). 1. Edition. Morgan Kauf-

mann.

Quillian, M.R. (1967). "Word concepts. A theory and simulation of some basic semantic capabilities". Behavioral Science 12 (5): 410430.

Van Dam, K.H. & Keirstead, J. (2010). Re-use of an ontology for modelling urban energy systems. In Proceedings of NGInfra 2010. Shenzhen.

J. Keirstead, N. Samsatli, A. M. Pantaleo, and N. Shah. Evaluating integrated urban biomass strategies for a UK eco-town. In Proceedings of the European Biomass Conference, pages 21152127, Hamburg, 2009.

K. H. van Dam. Capturing socio-technical systems with agent-based modelling. PhD thesis, Delft University of Technology, Delft, the Netherlands, 2009.

References

[1] BRE, BRE Housing Design Handbook, BRE, 1993. Report 253.