

Semantic Integration for Mapping the Underworld

Gaihua Fu and Anthony G Cohn

School of Computing, University of Leeds, Leeds, LS2 9JT, UK

Tel: +44 (0)113 3435430 Fax: +44 (0)113 3435468

{gaihua, agc}@comp.leeds.ac.uk

ABSTRACT

Utility infrastructure is vital to the daily life of modern society. As the vast majority of urban utility assets are buried underneath public roads, the need to install/repair utility assets often requires opening ground with busy traffic. Unfortunately, at present most excavation works are carried out without knowing exactly *what is where*, which causes far more street breakings than necessary. This research studies how maximum benefit can be gained from the existing knowledge of buried assets. The key challenge here is that utility data is heterogeneous, which arises due to different domain perceptions and varying data modelling practices. This research investigates factors which prevent utility knowledge from being fully exploited and suggests that integration techniques can be applied for reconciling semantic heterogeneity within the utility domain. In this paper we discuss the feasibility of a common utility ontology to describe underground assets, and present techniques for constructing a basic utility ontology in the form of a thesaurus. The paper also demonstrates how the utility thesaurus developed is employed as a shared ontology for mapping utility data. Experiments have been performed to evaluate the techniques proposed, and feedback from industrial partners is encouraging and shows that techniques work effectively with real world utility data.

Keywords: Urban Infrastructure, Semantic Heterogeneity, Data Integration, GIS

1. INTRODUCTION

Utility infrastructure, e.g. electricity and water systems etc., provides lifeline support for many aspects of modern society. As the vast majority of urban infrastructure is buried underneath public roads, the need to install or repair utility assets often requires opening ground with busy traffic. Unfortunately, at present most excavation works are carried out without knowing exactly *what is where*, due to the insufficient knowledge of buried services. This causes far more street breakings than would otherwise be necessary. According to [3], every year in excess of four million holes are dug in UK roads to repair/install assets, and the estimated cost of maintaining the nation's underground infrastructure is in excess of £3 billion per annum.

In order to avoid costly traffic delay and service disruption, many countries have legislation which requires that information of buried utilities must be obtained before any excavation occurs. However, the mapping information supplied by utilities is often of limited use. One main reason for this is that asset records are usually created and maintained by individual companies with little thought towards interoperability. This results in utility data differing from one company to another not only in *what* is encoded but also *how* it is encoded. This diversity of data resources makes it extremely difficult for the excavators to synthesize a homogeneous and integrated view of the excavation site, and results in unnecessary holes dug in the wrong place and third party damage to other underground services.

The exploitation of cost-effective techniques for the purposes of mapping underground utilities has been the subject of several studies. In some countries *One Call Systems* have been introduced [22]. They serve as a single contact point for excavators to determine which utilities are present in construction areas. One inadequacy of these systems is they identify/notify utilities that have services in the affected area, but do not attempt any integration or provide any enhanced value of utility data. Another stream of research looks at how utility records, usually residing on heterogeneous platforms, can be accessed through a common web portal [28]. The aim of these studies is to improve platform interoperability. The lack of effective ways for mapping utility data from multiple resources still remains a problem.

It is envisaged that more benefit could be gained from existing knowledge of buried utilities in street opening, if improved mechanisms of integrating knowledge could be provided. As a step towards this direction, the MTU and VISTA projects study heterogeneous data within the utility domain [18], and the ultimate aim of the research is to produce an integrated representation of underground utility infrastructure. The projects tackle a problem of great

practical significance and have already made a nationwide impact as well as attracting wide public and media interest. Strong links have been established with utility and other industries. More than 20 industrial companies contribute and participate in the research by amongst other things, providing utility data for the projects to test with and valuable feedbacks on our research results.

Both the MTU and VISTA projects research into techniques that reconcile syntax and semantic heterogeneity. This paper focuses on the techniques used to resolve the semantic heterogeneity. Semantics refers to *meaning* of data in contrast to syntax, which only defines the *structure* of data. Semantic heterogeneity of utility data arises due to different domain perceptions and varying data modelling and recording practices in the sector. Resolving this heterogeneity will remove domain inconsistencies and result in improved understanding of utility data. Semantic integration has been an active research area for the past few years [10, 14]. Despite efforts from the database and AI communities, semantic integration remains a difficult task. A principal reason for this is that the semantics of the data can be inferred from only a few available resources, thus integration results are often unreliable due to this incomplete information. The problem is more challenging in the utility domain due to a higher level of data heterogeneity, uncertainty and unreliability in the sector, and the lack of established utility data standards/ontologies which can be employed to support integration.

This paper investigates techniques for reconciling semantic heterogeneity within the utility domain. More specifically the research suggests the use of a shared knowledge structure (in form of utility thesaurus) to define the mapping among utility data. A bottom up approach was employed for the development of the thesaurus. The thesaurus provides a reference vocabulary on which to base the identification of mapping between utility data and subsequent resolution for semantic heterogeneity. Several experiments have been carried out to evaluate the techniques proposed. Usability and completeness of the ontology was evaluated by its ability to be mapped to utility data resources. Its effectiveness was evaluated by comparing it against existing utility data models/standards. Mapping results were sent to the relevant utility partners for validation, and statistics have been obtained to demonstrate the precision of the mapping. The experimental results we obtained are encouraging and show that techniques work effectively with real world utility data.

The remaining part of the paper is organized as following. Section 2 reviews related research. Section 3 presents techniques for thesaurus development. Section 4 describes how the thesaurus is employed to pinpoint utility data integration. Section 5 reports our experimental results. Section 6 concludes the paper and points out future research.

2. RELATED WORK

The heterogeneities of utility data are caused by many factors but the main reason is that utility data is autonomous, i.e. created and maintained by individual utility companies. Furthermore, the data is encoded in an uncoordinated way, i.e. without consideration of interoperability with other utility systems. For the purpose of discussion, we broadly classify heterogeneities into two categories: *syntactic* and *semantic* heterogeneity. Syntactic heterogeneity is concerned with the *structure* of data and semantic heterogeneity is concerned with the *meaning* of data. This paper is focusing on resolving semantic heterogeneity of utility data, and the reader is referred to [2] for our work on syntactic/schematic integration.

Semantics is the interpretation that people attribute to data (relating data to real world object). Different interpretations of data cause semantic heterogeneity. In [26], semantic heterogeneity is defined as the disagreement about the meaning, interpretation, or intended use of the same or related data – a definition also assumed throughout this paper. Semantic heterogeneity can occur at both schema level and data level. Examples of schema level heterogeneity include conflicts of class or attribute names, as well as the inconsistency on semantic constraints attached to schema level elements. Data level heterogeneity arises from differences in the data values returned by different databases for the same objects. A detailed discussion of semantic heterogeneities is presented in [13].

Overcoming semantic heterogeneity has been an active area of research in database and information integration communities. In past decades, a large body of research has been produced, ranging from mechanisms for accessing the data sources to techniques for matching database schema [4, 14]. Ontology research is another discipline that deals with semantic heterogeneity [24, 11]. While there are many definitions of what an ontology is, the common one is that ontology is some explicit specification of a domain of discourse, intended for exchanging and sharing data among different applications, and expressed in a language that can be used for reasoning. A distinctive feature of semantic integration research in the ontology community is that it enables integration to be performed by making use of the integration and reasoning tools developed in AI. In what follows, we will review previous research that is relevant to the research described in this paper. The discussion will be made on several areas, including integration architectures, mapping discovery methods, mapping representation. We refer the reader to [11, 14] for a thorough review of the area.

2.1 Integration Architectures

Systems developed to tackle semantic heterogeneities usually employ different system architectures. A *peer to peer* system (such as the ones described in [9, 1]) establishes direct semantic mapping among local data sources, and allows peers (i.e., participating data sources) to query and retrieve data directly from each other. All queries are made via a local vocabulary/ontology and semantic mapping among the peer data is used to compose local queries on other data sources. This approach looks more adaptive but requires the effort of creating mapping for each pair of local data sources.

Some systems are characterized by a global ontology (or data model) which represents a reconciled, integrated view of the underlying data sources. Notable systems of this type include the ones described in [23, 20]. Systems taking this approach usually provide users with a uniform interface -- all queries made to source data are expressed in terms of global ontology, thus freeing them from the need to understand each individual data source. The integration process requires establishing semantic mapping between the global ontology and each source data, and then using the mapping to answer queries. This approach looks straightforward but a shared ontology is required. Such an ontology can be constructed by integrating the local ones. A good practice is to reuse existing foundational ontologies. A number of very general ontologies that formalising events, time, space are being developed and some of them are becoming accepted standards. Unfortunately, little work has been done in literature to develop utility ontologies. In our investigations, the most up to date one is the utility data content standard proposed by the US Federal Geographic Data Committee [6]. Several projects have employed it as a shared data model for utility planning and other purposes [25]. However, the results of our utility mapping experiments (see Section 5) show that the knowledge encoded in FGDC standard is insufficient to serve as a reference model on which to base UK utility data integration.

2.2 Similarity Measure and Mapping Discovery

A fundamental operation for semantic integration is the similarity measure, which takes two or more data sources as input and produces a mapping between elements that semantically correspond to each other. Similarity measures are typically performed based on clues such as element names and integrity constraints. Almost all systems make use of element names in similarity measure. Some research goes a step further. For example, in [16], domain compatibility and referential constraints are exploited. In [21], *is-a* relationships and cardinality constraints are studied. Complex hierarchical relationships are exploited in [12] for a geo-spatial setting. Instance level information is examined in [15] to infer the schema matching information.

Based on the matching criteria proposed, various approaches have been suggested in the literature to find the mappings among data sources. Heuristic-based approaches employ hand-crafted rules [16] to find mapping. A common example is that two elements semantically match to each other if they have the same name and similar definitions. An alternative technique is to use learning based methods to discover matching pairs [4, 15]. For example, the research in [15] uses a neural network learning approach. It finds mappings based on attribute specifications (e.g. data types, scale, constraints etc.) and statistics of data content (e.g. maximum, minimum, average, and variance). The main benefit of learning-based approaches is that they maximally support automated integration, though efforts are required to obtain training data.

2.3 Mapping Representation

One of the most important aspects in semantic integration research is the specification of the mappings between matched elements. It is exactly this correspondence that will determine some fundamental problems in integration, for example, how the queries posed to the system are answered. With different interests in mind, several research communities provide a variety of support for this. The database community tends to use integrated views to describe mappings from a global schema to local schema. Two basic approaches have been proposed. The first approach, called global-as-view (GAV), requires that the global schema is expressed in terms of local ones [9]. The second approach, called local-as-view (LAV), requires the global schema to be specified independently from the sources, and the relationships between the global schema and the sources are established by defining every source as a view over the global schema [27]. In addition to GAV and LAV, other mapping approaches have been introduced such as GLAV [7] and BAV [19].

Ontology research tends to make use of the high expressive power of ontology languages to specify mapping [5, 24]. The main purpose of this is to enable complex mapping to be expressed, and translation between mapped elements to be enabled by utilising inference engines developed for ontology reasoning. For example, in [5] the correspondence between two ontologies is expressed as a set of bridging axioms relating classes and properties of two source ontologies. These axioms together with source ontologies are then treated as a single theory by a theorem prover optimised for

ontology translation task. Some research represents mapping as instances in an ontology which can then be used by an application to translate data from one source ontology to another [17].

3. THESAURUS DEVELOPMENT

An effective approach in semantic integration is to employ a shared ontology to base the identification of mapping between source data and subsequent resolution for semantic heterogeneity. Due to the lack of an accepted ontology in utility domain, this research develops a utility ontology to help with data integration. The initial development of the utility ontology takes the form of a thesaurus which maintains a controlled vocabulary describing utility asset feature types. Approaches in literature for building an ontology/ thesaurus can broadly be classified as one of two types: top-down approach or bottom-up approach. A top-down approach starts with the definition of the most general concepts in the domain and subsequent specialization of the concepts. A bottom-up approach starts with the definition of the specific concepts, with subsequent grouping of these classes into more general concepts. While neither of these two approaches is superior to the other, the general view is that the top-down approach is more suitable if the developers have thorough and good understanding of the domain, which is considered as difficult due to the inherent complexity of utility domains. In this research a bottom-up approach has been employed. The approach reuses and integrates the existing ontological utility categories and the overall process of thesaurus development consists of 6 steps, namely **Term Extraction**, **Relationship Derivation**, **Thesaurus Abstraction**, **Thesaurus Unification**, **Thesaurus Validation** and **Thesaurus Evaluation**. The process of the thesaurus development is iterative: revision and refinement may be performed at any stage to evolve the thesaurus, as illustrated in Figure 1. The thesaurus development tool **MultiTes Pro** is employed in our work for assisting thesaurus construction. In what follows, we will describe the first 5 steps of the thesaurus development, and we will elaborate step 6 in Section 5.

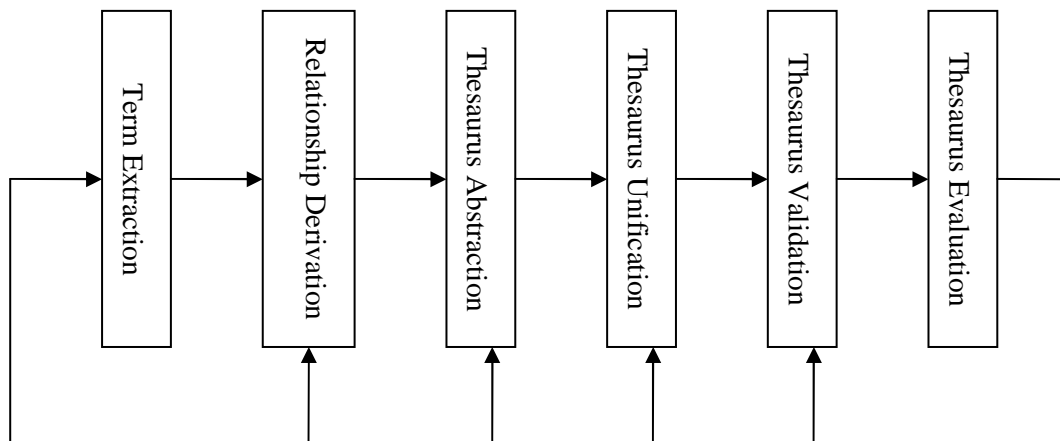


Figure 1 Thesaurus Development

Term Extraction

This step elicits from relevant resources the terms that are designed to specify utility asset types and subtypes, such as *water air valve* and *water hydrant*. The types of resources exploited here include the glossaries of utility terms which are available in research publications, text books, urban infrastructure design standards, and utility companies' web sites and so on. Another useful resource is utility data models and documents from utility companies, data standard organizations, and utility software vendors. Though this research mainly looks at written texts or documents for thesaurus construction, the input of domain experts has been playing an important role especially when ambiguity or uncertain arise.

Extracting terms from above documents is a straight forward process. For example, classes or entities are extracted from utility data models as thesaurus terms. Terms in utility glossaries are extracted as thesaurus terms. Both name and descriptive information (when available) of terms are extracted at this stage. The description of a term is denoted as **SN** (Scope Note). Each term is assigned to a subject category (**SC** for short), e.g. *water* or *sewer*.

Relationship Derivation

This step generates terminological relationships between utility terms. The following types of relationships are encoded in the thesaurus. Figure 2 shows an example thesaurus entry, which describes the term *water check valve*.

- **BT/NT** -- Broader Term/Narrower Term
- **USE/UF** -- USE/Used For
- **RT** -- Related Term (when terms are linked to each other through other semantic relationships other than BT/NT)
- **ST** -- Sibling Term (when terms are classified at same hierarchical level)

<p>Water Check Valve SC: Water SN: Devices that allow water to pass in one direction only. UF: Reflux Valve Water Backflow valve Water Flap Valve Water Non-ReturnValve BT: Water Valve</p>

Figure 2 An Example Thesaurus Entry

Two major sources for deriving terminological relationships are the utility data models and descriptive information of the utility terms (as exploited and acquired in the term extraction stage). Utility data models have been studied to identify several types of relationships. For example, **BT/NT** relationships are generated based on the inheritance relationships between classes/entities. **RT** relationships are generated by examining various associations between classes/entities.

Descriptions of thesaurus terms are another resource we examine to infer relationships, which are usually implicitly specified in one form or another. Clues for inferring these implicit relationships include the ways in which a term is described and meaning it delivers. For example, in some cases the descriptive information of a term **TERM1** is simply specified in form of “See **TERM2**” or includes a text like “... also called **TERM2**”. This usually allows us to infer a **USE/UF** relationship between **TERM1** and **TERM2**. Often, multiple relationships can be derived from a single piece of descriptive note. For example, the term *Abstraction Meter* may have following description “A type of water meters usually fitted at water sources”, which enables us to generate a **BT/NT** relationship between *Abstraction Meter* and *Water Meter*, and **RT** relationship between *Abstraction Meter* and *Water Source*.

Thesaurus Abstraction

This step groups utility terms in clusters in an iterative manner. That is, terms are broadly divided in big groups firstly according to their semantic similarity, and then each big group is sub-divided as smaller groups. For each group or sub-group, a term is singled out or introduced to abstract all terms in the group. In this way, the thesaurus is organised in a hierarchical manner and terms are allocated at different levels of granularity.

Thesaurus Unification

The **unification** step synthesises terms which specify the same feature types. For example, *building drain* and *house drain* can be unified as one term because they all refer to pipelines located inside a property for serving a customer. For each set of such terms, the unification step generates a single, consistent term definition which applies to all terms in the set. One term is singled out as the *preferred* term (based on the criteria that it is the one referred to most often by the utility companies or in the literature), and all other are *non preferred* terms. **USE/UF** relationships are established between preferred and non-preferred terms. Unification is performed for each term group generated in **abstraction** step. Usually no cross-group unification is required since typically terms which are semantically similar to each other are clustered in a same group.

Thesaurus Validation

This step deals with consistency checking of the thesaurus, which is mostly performed using thesaurus tool **MultiTes Pro**. **MultiTes Pro** maintains the consistency of a thesaurus in two ways. Firstly, it is able to detect any conflicts that an existing thesaurus exhibits. Secondly, it rejects adding any thesaurus terms or relationships that conflict with existing ones in the thesaurus. **MultiTes Pro** can also infer implicit relationships based on the nature of a terminological relationship. For example, it is able to infer that **TERM2 is a NT term of TERM1** from the fact that **TERM1 is a BT term of TERM2**.

4. MAPPING DISCOVERY

The utility thesaurus developed is employed as a shared ontology on which to base the identification of mapping between utility asset type and subtype codes (utility codes for short). Such codes have been used by utilities to specify various asset features stored in their utility databases, and they are usually embodied themselves as schema names or data instances. The essential first step of data integration is to extract utility codes and their interpretation from available resources, e.g. database tables, lookup definitions and other meta-data resources. The results of this extraction include the code name, code definitions, and information about where codes are defined and where codes are used.

Mapping is the task of relating utility codes of two or more utility resources. The generic utility thesaurus constructed has been used as a reference ontology to support this. The assumption is that if utility codes coming from different utilities can be matched to a same thesaurus term, then these utility codes are equivalent (or similar in approximate matching), and they refer to the same (or similar) real world objects. Figure 3 shows how utility asset types are encoded differently by different companies, and how data from different companies can be matched to each other by grounding to the generic utility thesaurus. We use different colours/shades here to distinguish codes from different data resources. Utility asset type codes linked to the same term are semantic equivalent or similar.

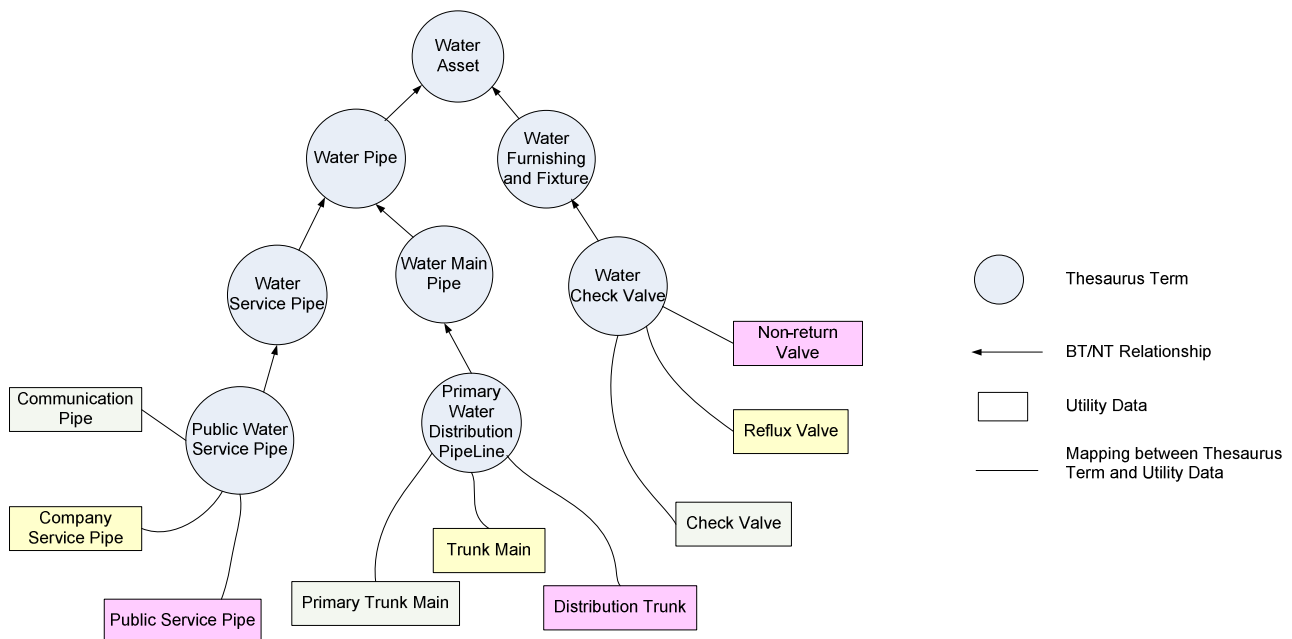


Figure 3 Mapping Utility Data to Thesaurus

The mapping discovery relies on semantic matching heuristics to find the matches. Similarity of two elements is judged based on the characteristics of the specified elements and similarity of their related elements. The main information exploited here are utility code/thesaurus term names, definitions or descriptions codes/terms, the relationships of specified codes/terms with other codes/terms.

Each round of the mapping process takes the thesaurus and codes from one utility company as input, and establishes a set of relations between utility codes and thesaurus terms. Mapping for all participating companies comprises the complete mapping. For each utility code, the mapping process generates a set of candidate thesaurus terms and identifies the one which best matches the code. Sometimes domain experts from relevant utility companies may be asked to provide some inputs to help with decision making. A match result is identified as one of two types: *exact match* or *approximate match*. An exact match finds a thesaurus term which is semantically *equivalent* to the specified utility code. When an exact match can not be made, mapping performs an approximate match – finding a term that is semantically most similar. For example, consider the code *Double Orifice Washout Hydrant*: if the thesaurus does not have an equivalent term, it may be matched to the term *Washout Hydrant* if it is the most semantically similar.

The mapping process identifies several types of approximate matches, namely *BT match*, *NT match*, *ST match*. BT match maps a utility code to a term that is semantically *broader* than it, as illustrated by the above example. NT match maps a code to a term that is semantically *narrower* than it. ST match maps a code to a term that is semantically a *sibling* to it. Figure 4 shows several possible approximate matches for utility code *Water Tower* when an exact match can not be made for it. While the approximate mapping may potentially generate different approximate matches, we consider that a *BT match* is a more desirable resolution than other approximate match types. The underlying rationale is that a broader term *subsumes* its narrower terms (an instance of a narrow term is always an instance of its broader term), and therefore matching a utility code to a broader term still correctly (though not completely) preserves its semantics. This is not the case with a NT or a ST match where slightly different data semantics may be carried by a NT or ST term. However, when an immediate BT term is not available, we consider mapping a utility code to an immediate NT or a ST term may be more meaningful. For example, for the example shown in Figure 4, if the term *water storage* is not available, then matching the code *water tower* to the term *water Tank* is more meaningful than to *Water Furnishing and Fixture*¹.

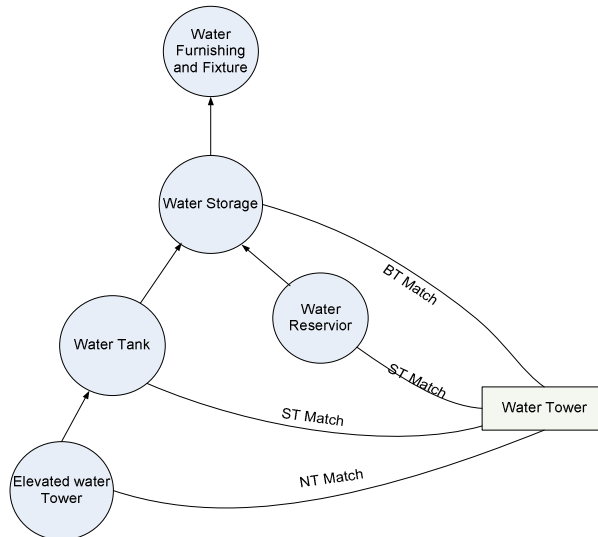


Figure 4 Approximate Mapping

As a utility code may match to a thesaurus term in different ways, codes grounded to the same thesaurus term may have complex mapping relationships to each other. Appendix A shows a set of axioms that can be used to deduce the mapping relationships between two utility codes C_i and C_j when they both match to a same thesaurus term T .

Several observations have been made in our mapping exercises. Firstly, information available for performing mapping varies from one utility company to another, and therefore the matching heuristic may need to be adjusted to fit for the purpose of each company. Secondly, utilities usually organise asset types in flat structures – features with different granularities and categories are often clustered together. Mapping utility codes to a shared thesaurus enables us to gain a clear and better understanding of the utility asset hierarchy. However, research still is required on types of mapping best suited to potential users such as street workers (e.g. in some case a BT match may be more appropriate than an exact

¹ We are still in the process of designing techniques to compute weighted scores for different match types, and we hope to report this result very soon.

mapping). Thirdly, matching experiments do not always result in *1-1* mapping, i.e. each utility code uniquely assigned to one thesaurus term. Due to the semantic mixture or semantic fragment, sometimes *1-n*, *n-1* or *m-n* mapping may exist. Finally, it is not always possible to perform a *total* mapping, i.e. not every code has a match in the thesaurus (this is called a *partial* mapping). There are two main reasons for this. One is the completeness of the thesaurus, i.e. it may not completely conceptualise the domain in its initial development, and therefore it needs to evolve based on the mapping results. The second is due to nature of utility codes e.g. some codes may be introduced by utilities to encode abstract asset objects, and thus they are deemed to have no matches.

5. EXPERIMENTS AND EVALUATION

To evaluate the techniques proposed, a thesaurus has been constructed for the water domain. A total of 248 terms have been captured in the initial development of the thesaurus, and are structured in 6 levels of hierarchy². Utility codes from several water companies were matched to thesaurus. Mapping results was sent to the relevant utility companies for validation. We report some of experimental results in the remaining part of this section.

5.1 Mapping Evaluation

Utility codes from four UK water companies (anonymous here for reasons of confidentiality) have been collected and matched to the thesaurus. Figure 5 shows the total number of codes collected for each company and number of codes with matches from the thesaurus. Though the percentage of matches varies from one company to another, the overall match percentage is satisfactory – on average about 94% percent of codes are mapped to thesaurus terms. The manual checking of non-matching codes reveals that the lack of interpretation of utility codes is one main reason for not having found matches for those codes. Another reason is that many of these codes are designed to describe abstract asset types and therefore can not be matched to the thesaurus which was designed only to capture physical asset types.

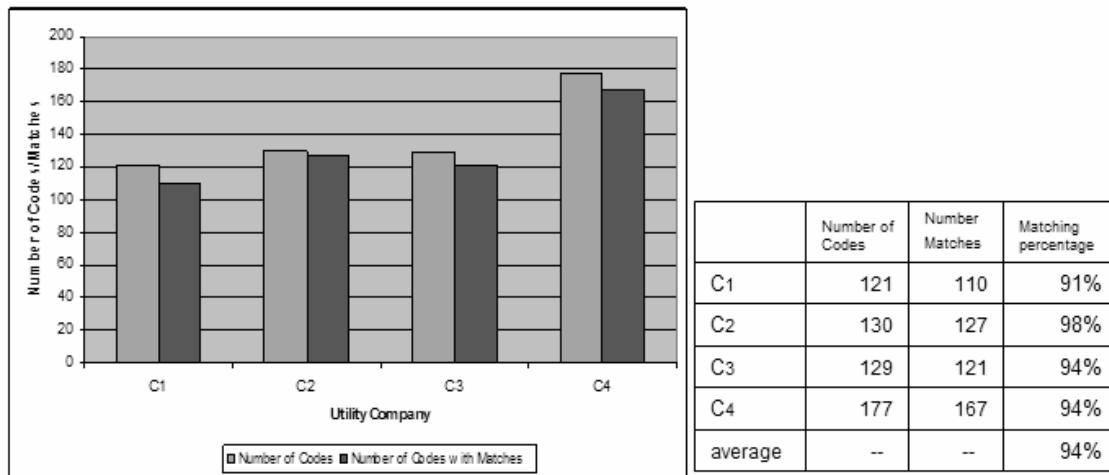


Figure 5 Mapping Utility Codes to the Thesaurus

The mapping identifies both *exact* and *approximate* matches. The bar chart in Figure 6 illustrates the number and percentage of these two types of match, as well as the number and percentage for non-matched codes. The figure shows that the majority of matches found were exact matches (around 80% on average). About 14% were approximate matches, and our investigation revealed that they were BT matches exclusively, which is the most desirable solution for approximate matches as pointed out in Section 4. This implies that although the thesaurus developed so far does not fully cover the semantics of some utility codes, the availability of these broader terms allows us to easily extend the thesaurus (without changing its structure) to accommodate these missing terms.

The mapping results have been sent to utility companies for validation. For each mapping pair, utility companies were asked to select from one of three options which best described the match: *correct match*, *incorrect match*, *unsure match*.

² The thesaurus evolves to accommodate any missing terms or other data semantics as identified in the mapping.

They were asked to provide further comments for codes that were not matched to a thesaurus term. Utility companies were also asked to provide an explanation or comment if they classified a match as *incorrect* or *unsure*. This information will be employed to improve the quality of the thesaurus and help with thesaurus evolution. Table 2 summarises the feedback we have received so far, where N is the number of matches returned by our mapping experiments, M is the number of mapping records evaluated by utility companies, N1 is the number of correct matches identified, N2 is the number of incorrect matches identified, N3 is number of unsure matches, and N4 is number of comments we received.

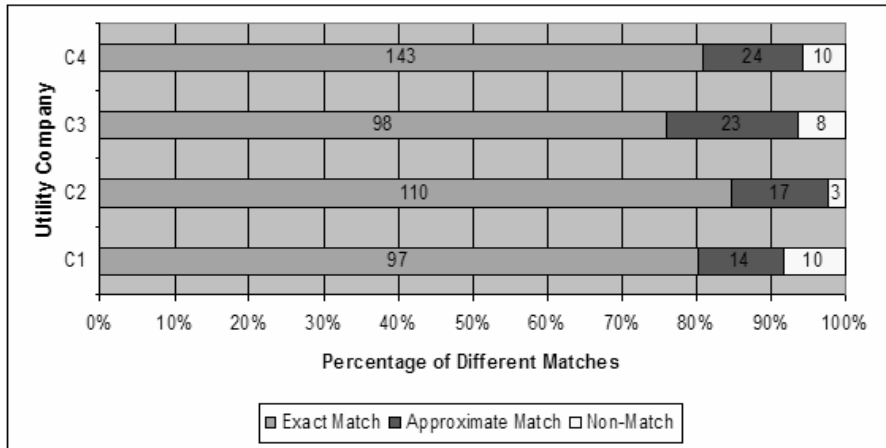


Figure 6 Distribution of Matches

We received varying levels of feedback from the utility companies. Based on the feedback we received, we calculated the precision of mapping, which is shown in the last column of Table 1. The mapping precision P is defined here as the percentage of correct mappings identified by utility domain experts and returned by our mapping experiments. As not all utilities performed a full evaluation, the precision was calculated as $P=N1/M$. From Table 1 we can see that mapping precision differs from one company to another. On average, the mapping precision is satisfactory. As the thesaurus development is iterative – it evolves to respond to mapping experiments, it is anticipated that the mapping precision will gradually improve in the future mapping experiments.

Table 1 Mapping Evaluation by Utility Companies

	N	M	N1	N2	N3	N4	P
C1	110	-	-	-	-	18	-
C2	127	90	82	7	1	15	91.1%
C3	120	120	104	11	5	56	86.7%
C4	162	162	146	15	1	39	90.1%
average	-	-	-	-	-	-	89.3%

5.2 Ontology Evaluation

The experiments described in Section 5.1 demonstrated that the developed thesaurus has a good coverage of the domain – on average about 94% percent of codes have matches in the thesaurus. The effectiveness of thesaurus was further demonstrated by comparing it to the FGDC utility data standard. For the purpose of this study, we extracted from FGDC a set of water asset terms, and constructed a thesaurus using the information acquired. Table 2 shows a general comparison of FGDC thesaurus and the thesaurus developed in this research (MTU thesaurus for short). The two thesauri capture similar numbers of terms. The FGDC thesaurus has a quite flat structure, and terms are organised in 2 levels of hierarchy. The main drawback for this is that asset types of different granularity are clustered in same group. The MTU thesaurus overcomes this by allocating asset types to appropriate levels of hierarchies, and currently 6 levels of hierarchy are maintained. In term of semantic relationships, we did not find that the FGDC encodes multiple-inheritance

relationships between terms, which are effectively supported by the MTU thesaurus. Finally, the MTU thesaurus supports a richer set of semantic relationships than does the FGDC as illustrated in Table 2.

Table 2 Structural Comparison of MTU Thesaurus and FGDC Thesaurus

	MTU Thesaurus	FGDC Thesaurus
Number of terms	248	222
Level of hierarchies	6	2
Multi-inheritance allowed?	yes	no
Relationships supported	BT, NT, SN, SC, ST, USE, UF, RT, ST	BT,NT,SN,SC

To compare how well the two thesauri serve the mapping purpose, we matched same set of utility code to the FGDC thesaurus and compared results with the MTU thesaurus mapping, which is shown in Figure 7. The results reveal that when used as a shared thesaurus to integrate utility codes, the MTU thesaurus out-performed the FGDC thesaurus both on a per utility company basis but also on average overall: only 78% codes found matches with the FGDC thesaurus, which was 16% less than MTU thesaurus mapping. Further investigation revealed that approximate matches counted for 52% of all matches with FGDC thesaurus mapping. Among all approximate matches, around 83% were BT matches, 8% were NT matches, and 9% were ST matches. This is significantly different from the approximate matches made to the MTU thesaurus, where approximate matches (which were BT matches exclusively) counted for 14%.

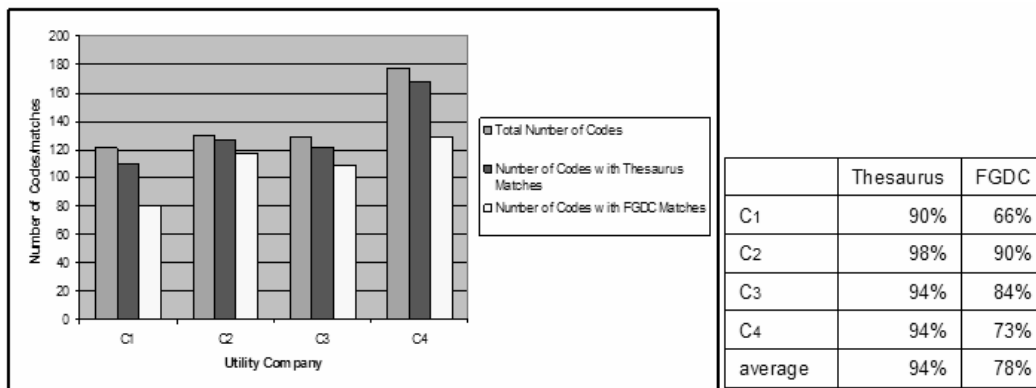


Figure 7 Mapping Comparison between the MTU Thesaurus and the FGDC Thesaurus

5.3 Heterogeneity Study

The mapping experiments provided a good test bed to investigate and quantify the scale of the heterogeneity existing with utility data. Utility companies vary from each other on how they conceptualise the domain, which causes data heterogeneity. By mapping utility codes to a single thesaurus, we were able to obtain statistics on the heterogeneity of utility codes. The first study was on inter-utility heterogeneities. The investigation was made on the common asset types captured by asset recording systems of different companies. Our statistics revealed that the heterogeneity of utility codes on this is significant. As shown in Figure 8, among all thesaurus terms matched to utility codes, only 9% of terms have matches from all 4 companies, 11% from 3 companies, 23% from 2 companies and with large majority 57% matched only to 1 company. This suggests that our underlying assumption – inter-utility heterogeneities widely exist – is correct.

Heterogeneities not only exist cross utilities, but also within a single utility company, i.e. the same asset types are specified with different codes. This intra-utility heterogeneity often results in that same number of utility codes does not exactly match to the same number of thesaurus terms. For example, one utility company involved in the mapping experiments has 129 codes, and 121 of them have matches and these were only matched to 80 thesaurus terms. Due to the space limitation, we are not able to present further details here.

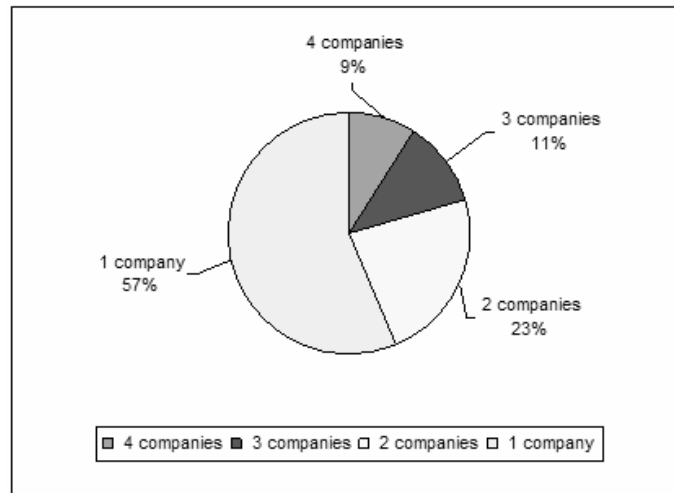


Figure 8 Common Asset Types Encoded by Different Companies

6. CONCLUSIONS

This paper presents our research on semantic integration of utility data for the purpose of providing an improved information mapping service for underground utility apparatus records. The research develops techniques for constructing a basic utility ontology, and also techniques for mapping utility data to the ontology developed. Experiments have been performed to evaluate the techniques proposed, and feedback from industrial partners is encouraging and shows that techniques work effectively with real world utility data. The techniques add several benefits to a utility mapping service, the key one being that it removes domain inconsistencies and results in improved understanding by employing transparent naming and representation standards. We consider that semantic integration is one important technological aspect of a fully successful approach to managing, sharing and making best possible use of utility sector data. There are a number of areas which we plan to explore in the future, including development of techniques to extend the ontology to capture richer data semantics, in order to bridge the gap between the schematic and semantic integration of utility data. We also plan to construct an ontology that covers different utility domains.

ACKNOWLEDGEMENT

This work is funded by EPSRC Grant EP/C014707/1 and DTI Grant DTI/TSB 15820. Thanks are given to our utility partners for their time and effort for evaluation the mapping results, as well as the valuable domain knowledge they provided. Thanks are also due to Jo parker from UKWIR for her effort of coordinating the mapping evaluation work.

REFERENCES

- ¹ Y. Arens, C. A. Knoblock, C. Hsu: Query Processing in the SIMS Information Mediator, in *Advanced Planning Technology*, A. Tate (ed), AAAI Press, 1996.
- ² A. Beck, A. G. Cohn, M. Sanderson, S. Ramage, C. Tagg, G. Fu, B. Bennett, J. Stell, UK Utility Data Integration: Overcoming Schematic Heterogeneity, extended abstract to appear in *GeoInformatics2008*.
- ³ M. Burtwell, E. Faragher, D. Neville, C. Overton, C. Rogers, and T. Woodward. Locating Underground Plant and Equipment: Proposals For a Research Programme. *Technical Report TR03-wm-12-4, UKWIR*, 2004.
- ⁴ A. Doan, P. Domingos, and A.Y. Halevy, (2001). Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach. *SIGMOD record*, 30(2), 509-520.
- ⁵ D. Dou, D. V. McDermott, and P. Qi. Ontology Translation on the Semantic Web. In *OTM Confederated International Conferences, CoopIS, DOA, and ODBASE*, pages 952-969, 2003.
- ⁶ Federal Geographic Data Committee, Utilities Data Content Standard. <http://www.fgdc.gov/standards/projects/FGDC-standards-Projects/utilities/utilities.pdf>.

7 M. Friedman, A. Y Levy and T. D. Millstein *Navigational Plans For Data Integration*. in *Proceedings of the 16th National Conference on Artificial Intelligence*, 1999.

8 A.Y. Halevy, Answering Queries Using Views: A Survey. *The VLDB Journal*, 10(4), 270-294, 2001.

9 A. Y. Halevy, Z. Ives, D. Suci and I. Tatarinov, *Schema Mediation in Peer Data Management Systems*. In *Proceedings of the 19th International Conference on Data Engineering*, 2003.

10 R. Hull. Managing semantic heterogeneity in databases: a theoretical prospective. In *Proceedings of ACM symposium on Principles of Databases*, pages 51–61, 1997.

11 Y. Kalfoglou and M. Schorlemmer, Ontology Mapping: The State of the Art, in *Proceedings of Dagstuhl Seminar - Semantic Interoperability and Integration*, 2005.

12 M. Kavouras and M Kokla, A method for the formalization and integration of geographical categorizations, in *International Journal of Geographical Information Science* 16(5), page 439 – 453, 2002.

13 W. Kim and Jungyun Seo, Classifying Schematic and Data Heterogeneity in Multidatabase Systems, in *Computer*, 24 (12), pages 12-18, 1991.

14 M Lenzerini. Data Integration: A Theoretical Perspective. In *Proc. PODS 2002*, 2002.

15 W.S. Li and C. Clifton, SEMINT: a Tool For Identifying Attribute Correspondences in Heterogeneous Databases Using Neural Networks. In *Data and Knowledge Engineering*, 33(1), 49-84, 2000.

16 J. Madhavan, P. Bernstein and E. Rahm., Generic schema matching with cupid. In *The VLDB Journal*, 49-58, 2001.

17 A. Maedche, B. Motik, N. Silva, and R. Volz. MAFRA - An Ontology Mapping Framework in the Semantic Web. In *Proceedings of the ECAI Workshop on Knowledge Transformation*, 2002.

18 Mapping the Underworld Project, <http://comp.leeds.ac.uk/MTU>.

19 P. McBrien, P and A. Poulouvasilis, Data integration by bi-directional schema transformation rules. In *Proceedings of the 19th International Conference on Data Engineering*, 2003.

20 E. Mena, V. Kashyap, A. P. Sheth, A. Illarramendi: OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. In *Proceedings of the First IFCIS International Conference on Cooperative Information Systems*, 14-25, 1996.

21 P. Mitra, G. Wiederhold, and M. Kersten. A Graph-Oriented Model for Articulation of Ontology Interdependencies. In *Proceedings of international Conference on Extending Database Technology*, page 86-100, 2000.

22 Moleseye. One Call UK Online. <http://www.moleseye.net/Homepages/index.asp>.

23 A. Motro, J. Berlin and P. Anokhin. Multiplex, Fusionplex, and Autoplex - Three Generations of Information Integration. *SIGMOD record*, 33(4), 51-57, 2004.

24 N. F. Noy, Semantic Integration: a Survey of Ontology-Based Approaches, in *SIGMOD Record*, 33(4), page 65-70, 2004.

25 H. Osman and El-Diraby, Ontological Modeling of Infrastructure Products and Related Concepts, in *Transportation Research Record* (In Press).

26 A. P. Sheth and J.A. Larson, Federated Database Systems for Managing Distributed, Heterogenous, and Autonomous Databases. In *ACM Computing Surveys* Vol 22(3), page 183-236.

27 J. D. Ullman. Information Integration Using Logical Views. *Theoretical Computer Science*, 239(2):189-210, 2000.

28 YEDL system access. <http://www.safedig.co.uk>.

Appendix A Axioms for Deducing Relationships between Utility Codes

- $Axiom_1$ $ExactMatch(C_i, T) \wedge ExactMatch(C_j, T) \Rightarrow equivalence(C_i, C_j)$
- $Axiom_2$ $ExactMatch(C_i, T) \wedge NTMatch(C_j, T) \Rightarrow NT(C_i, C_j)$
- $Axiom_3$ $ExactMatch(C_i, T) \wedge BTMatch(C_j, T) \Rightarrow BT(C_i, C_j)$
- $Axiom_4$ $ExactMatch(C_i, T) \wedge STMatch(C_j, T) \Rightarrow ST(C_i, C_j)$
- $Axiom_5$ $NTMatch(C_i, T) \wedge NTMatch(C_j, T) \Rightarrow ST(C_i, C_j) \vee equivalence(C_i, C_j)$
- $Axiom_6$ $NTMatch(C_i, T) \wedge BTMatch(C_j, T) \Rightarrow BT(C_i, C_j)$
- $Axiom_7$ $NTMatch(C_i, T) \wedge STMatch(C_j, T) \Rightarrow BT(C_i, C_j)$
- $Axiom_8$ $BTMatch(C_i, T) \wedge BTMatch(C_j, T) \Rightarrow ST(C_i, C_j) \vee equivalence(C_i, C_j)$
- $Axiom_9$ $BTMatch(C_i, T) \wedge STMatch(C_j, T) \Rightarrow BT(C_j, C_i) \vee unrelated(C_i, C_j)$
- $Axiom_{10}$ $STMatch(C_i, T) \wedge STMatch(C_j, T) \Rightarrow ST(C_i, C_j)$