

SEEING THE UNSEEN: DELIVERING INTEGRATED UNDERGROUND UTILITY DATA IN THE UK

Anthony R. Beck^{a*}, Anthony G. Cohn^a, Jo Parker^b, Nadia Boukhelifa^a, Gaihua Fu^a

^a School of Computing, University of Leeds, Leeds, LS2 9JT, UK - (arb, agc, nadiab, gaihua)@comp.leeds.ac.uk

^b UK Water Industry Research (UKWIR) Ltd., 1 Queen Anne's gate, London SW1H 9BT, UK - j.parker101@btinternet.com

KEY WORDS: Utility, spatial, integration, GIS, visualization, heterogeneity, semantic, schematic, syntactic

ABSTRACT:

In earlier work we proposed a framework to integrate heterogeneous geospatial utility data in the UK. This paper provides an update on the techniques used to resolve semantic and schematic heterogeneities in the UK utility domain. Approaches for data delivery are discussed, including descriptions of three pilot projects and domain specific visualization issues are considered. A number of practical considerations are discussed that will impact on how any implementation architecture is derived from the integration framework. Considerations of stability, security, currency, operational impact and response time can reveal a number of conflicting constraints. The impacts of these constraints are discussed in respect of either a virtual or materialised delivery system.

1. INTRODUCTION

Utility companies supply essential services to industry and homes throughout the world. Collateral information about other utility services in the same locality is essential for the safe, effective and efficient management and maintenance of utility assets. This is particularly important for those assets which are buried and therefore unseen. In addition there are a range of different third party users who pay for access to utility data. Many of these are for utility enquiries that underpin transactions in different market segments (for example conveyancing, insurance assessments, environmental impact assessments and planning applications). Furthermore, the effectively invisible underground assets owned by utility companies represent a significant health and safety hazard and mitigation problem to construction companies. This is particularly relevant to those working on brownfield sites or conducting major urban infrastructure projects (such as the Olympic village in London). Access to digital integrated utility data at the planning and construction stages can significantly reduce delays and improve health and safety conditions.

In the UK there are a range of different utility companies supplying services necessary to modern urbanism; gas, petrochemical, sewer, water, electricity and telecoms. Each company has a different approach to managing their asset network, although most store their data digitally and articulate it using Geographical Information Systems (GIS). The quality of the digital data can be variable: recently installed assets may have been well mapped, although, location and attribute data on older services can be very poor. Furthermore, the spatial inaccuracies of these data are unknown. For example, a utility company may be confident that it knows where 90% of its assets are, to a certain accuracy specification, but does not know where the 10% of unrecorded assets exist in its network. Marvin and Slater (1997) estimate that the location of only 50% of buried infrastructure is accurately known. In addition there is variable information pertaining to the third dimension (depth or elevation). Until recently only the sewer domain collected and stored depth and elevation data with any rigour. This is likely to change with the regularly use of GPS surveys. This variability in data quality can lead to uncertainty.

Prior to invasive works it is normally required that excavators should request and obtain record information from all relevant utilities to identify what is buried where. The current industry approach is based upon requesting paper based output, which can take weeks for delivery, or downloading raster outputs from a webGIS. Both of these approaches generalise the data attributes and require conversion to integrate the data with digital spatial decision support tools. Unfortunately, external companies (including competitor utilities, construction projects and highways authorities) tend to have a low level of access to the rich data held at source which results in a dilution of knowledge about the asset (NUAG, 2006). Some information held in utility records, e.g. installation details, maintenance history and physical properties of buried assets which are relevant to excavation works are not articulated. Knowledge sharing is hindered as each utility company employs their own methods for data recording and presentation and there is significant variability within each sector.

In summary the majority of utility companies have robust systems for managing their assets and facilities. However, the data can be imperfectly populated and has a range of spatial and attribute errors. The current approaches to sharing this data are time consuming and/or dilute the information resource. The core of the problem is that data consumers have insufficient and inadequate knowledge about what is where concerning buried assets. Furthermore, the available knowledge is not accessible in a timely manner. This is time, cost and process inefficient for utility companies (statutory undertakers) and those third party organisations that rely on utility data in their business.

We postulate that improving mechanisms of integrating and sharing knowledge on utility assets and the location of street works will lead to a reduction in the amount of works through better co-ordination and information quality. It is important to note that by quality we mean the modes and mechanisms in which information is shared and the harmonisation of structure and semantics as opposed to the underlying data quality.

This paper draws on the research undertaken by the School of Computing at the University of Leeds in both the Mapping The Underworld (MTU) and Visualizing integrated information on

buried assets to reduce streetworks (VISTA) projects (www.comp.leeds.ac.uk/mtu). VISTA commenced in early 2006 and is a four year joint project with a consortium of academic and industry partners. UK water industry research (UKWIR) is the lead co-ordinating partner with Leeds and Nottingham Universities providing the research input. In addition, there are over 20 utility and other industrial organisations. VISTA builds on the pre-existing Engineering and Physical Sciences Research Council funded MTU project.

These projects have developed technology to enable underground asset data from multiple information sources to be integrated in a common format. The research is motivated in response to the market need for better and more timely access to information pertaining to underground utility apparatus. Although both projects have a UK focus, the problem has worldwide applicability. This paper discusses the nature of utility heterogeneity and the proposed integration framework which have both been discussed in previous publications. We then discuss techniques for delivering integrated data, the preliminary results of three pilot projects and potential implementation considerations.

2. UTILITY HETEROGENEITY

Each utility company has developed its own approach for managing the digital view of their asset network. The design is based upon each company's abstracted view of their infrastructure and their business model. Hence, when analysed at the sector or domain level, data is encoded in an uncoordinated way, without consideration of compatibility and interoperability with other utility systems. Overcoming these heterogeneities is an essential first step in achieving utility integration and a move towards domain interoperability.

In common with other organisations that hold geospatial data, heterogeneities in the utility domain can be broadly grouped into the three categories discussed by Bishr (1998): syntactic, schematic and semantic heterogeneity. Beck *et al.* (2008) discuss the range of heterogeneities expressed in the UK utility domain in greater detail.

Syntactic heterogeneity refers to the difference in data format. The most profound difference is in the storage paradigm: relational or object orientated. Partner utility companies rely on a range of storage solutions including Oracle, SQL server and ArcSDE. However, some utility companies are starting to make their data available in Open Geospatial Consortium (OGC) approved syntactically interoperable formats and services such as Geography Markup Language (GML) and Web Feature Service (WFS). Syntactically interoperable approaches underpin a number of geospatial integration frameworks currently under development based on Service Oriented Architectures (Donaubauer *et al.*, 2007; Klien *et al.*, 2007; Lemmens *et al.*, 2007).

Schematic heterogeneity refers to the differences in data model between organisations. The database schema is designed at the conceptual modelling stage and reflects each company's abstracted view of their business and physical assets. Hence, different hierarchical and classification concepts are adopted by each company to refer to identical or similar real world objects.

Semantic heterogeneity refers to differences in naming conventions and conceptual groupings in different organisations. Naming mismatch arises when semantically

identical data items are named differently or semantically different data items are named identically. Naming heterogeneities can be relatively easily reconciled with a thesaurus. Different companies, or utility domains, have subtly different cognitive views of the world which means that they describe similar real world objects from different perspectives. Cognitive semantics can be subtle, reflecting the domain of discourse. For example, a road is seen by the traffic management community as a link in a topological transportation network whereas in the utility industry it is seen as a surface with different engineering properties, reinstatement issues and access constraints (Aerts *et al.*, 2006). Reconciling these cognitive heterogeneities is more difficult but is achievable through ontology mapping.

There are other utility domain specific heterogeneities that remain to be resolved. For example, different units and reference systems are a problem, although, this is reasonably constrained as all partner companies use the Ordnance Survey (OS) National Grid projection. However, the Positional Accuracy Improvement (PAI (OS, 2007)) programme, used to address accuracy issues in OS data that became apparent after using absolute positioning technologies (such as GPS), provides a 95% accuracy estimate of 1m in urban environments. The differences in precision and accuracy of relative and absolute positioning devices may increase data uncertainty.

Finally, though the literature is rich on techniques for resolving various heterogeneities, the assumption is that various metadata and documentation is available to assist integration work. This metadata is not always available: without good quality metadata some problems may be intractable (Bernard *et al.*, 2003).

3. A FRAMEWORK FOR DATA INTEGRATION

The previous sections have introduced the nature of utility asset data in the UK and the range of heterogeneities that exist within the utility domain. In response to this, we have designed a conceptual framework which supports utility knowledge and data integration. This framework was reported by Beck *et al.* (2008), what follows is a summary of the salient points of the framework and a description of the current progress in resolving syntactic, schematic and semantic heterogeneities.

The framework is characterised by a number of features:

- The framework supports utility integration at two levels: the schema level and the data instance level. The schema level integration ensures that a single, unified interface is provided to access utility data in a consistent way, and to enable underground asset data from multiple utilities to be represented in a common format. The data level integration improves utility data quality by reducing semantic inconsistency, duplication and conflicts.
- A global schema (common data model) based architecture is adopted.
- A bottom up approach is employed to construct the global schema/model of utility data. This contrasts with many other domains, where shared, standard models/ontologies exist (Fu and Cohn, 2008a; 2008b).

As shown in Figure 1, the main components of the framework are the Schema Integration Manager, Data Integration Manager and Query Manager. The Schema Integration Manager is designed to support schema level integration, which is mostly

performed as a pre-processing stage. It takes as input schema level knowledge, government legislation, codes of practice and users' knowledge as inputs, and produces the global schema mapping metadata between global and local schemas.

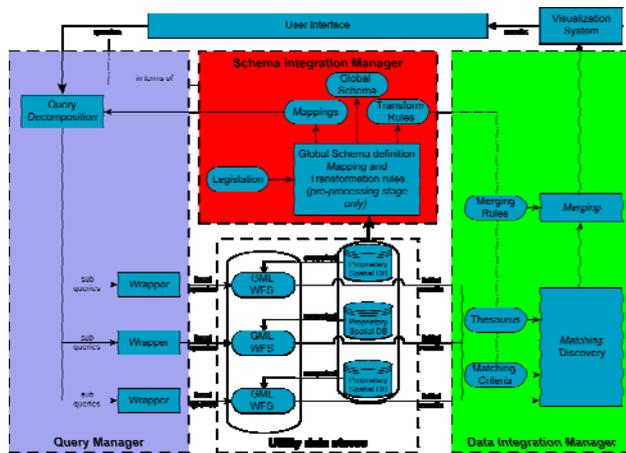


Figure 1, The proposed VISTA Framework for Utility Knowledge and Data Integration

The Data Integration Manager together with the Query Manager support run time integration activities. The query manager is necessary if a virtual integration approach is taken. All queries are specified in terms of the global schema. A query submitted to the utility integration system is first sent to the Query Manager. Based on the mappings generated by the Schema Integration Manager, the Query Manager decomposes the query into several local queries specified in terms of the underlying local schemas. These local queries are then forwarded to the appropriate utility datastore. These datastores are snapshots of the operational datastores. As they will form part of the business process for each utility company their refresh rates will be controlled by each individual organisation. The query results produced by the local datastores are integrated in the Data Integration Manager and then sent back to the user interface via a visualization system.

3.1 Syntactic Integration

It is worthwhile noting that during the proof of concept phases all data is held within an Oracle Spatial repository. Extract, Transform and Load (ETL) middleware was used to convert the range of supplied data into an Oracle Spatial format.

In a production environment it is envisaged that data would be accessed using an interoperable OGC compliant approach. When utility data was requested from the industrial partners it was asked if it could be provided using OGC compliant mechanisms. It was interesting to note that by early 2008 most partner organisation were able to produce GML files or were implementing WFS and cited the UK Traffic Management Act (NUAG, 2006) as one of the drivers.

3.2 Semantic Integration

Semantic heterogeneity can be resolved through the use of an ontology or thesaurus. Both techniques define a common vocabulary to reduce the semantic problems relating to data interoperability and integration. However, an ontology or thesauri does not in itself guarantee interoperability (Cai, 2007).

A thesaurus comprises of a collection of significant terms and the relationships between those terms in a domain knowledge space. A thesaurus has explicit concepts, relations, terminology and definitions (explicit specification) and reflects a consensual abstracted knowledge model for a domain (shared conceptualisation: Bernard et al., 2003; Deliiska, 2007; Studer et al., 1998).

An ontology is defined as a formal, explicit specification of a shared conceptualisation (Gruber, 1993). Hence, an ontology is essentially a machine readable (formal) thesaurus. Therefore an ontology allows sophisticated information processing, such as inferred properties, based on logical reasoning (Klien et al., 2007). For this reason an ontology is preferred over a thesauri approach. However, these are not competing techniques as an ontology is a richer conceptualisation of a thesaurus.

The VISTA project has built a thesaurus for the sewer, water and electricity utility domains using the 'MultiTes Pro' software (www.multites.com: Fu and Cohn, 2008a). The thesaurus maintains a controlled vocabulary describing utility asset feature types and related terms. The overall process of thesaurus development consists of 6 steps, namely Term Extraction, Relationship Derivation, Thesaurus Abstraction, Thesaurus Unification, Thesaurus Validation and Thesaurus Evaluation (cf. Ahmed et al., 2007; Deliiska, 2007; Pinto and Martins, 2004; Pundt, 2007). In addition Formal Concept Analysis has also been used for thesauri development (Fu and Cohn, 2008b). The resultant thesaurus has been used as a shared vocabulary to map utility asset type and subtype data.

The utility thesaurus acts as a shared vocabulary that is used to reconcile data value heterogeneities for fields in the Global Schema (including asset subtype, service type, and material type values: see Table 1). Mapping discovery, based on matching heuristics, identifies the relationship between the source utility value and the thesaurus terms. Values mapped to the same thesaurus term are considered to be equivalent or similar. The main information exploited here are value names, value definitions or descriptions and the relationship of one value to another. The mapping experiments were performed on 5 water datasets, 4 sewerage datasets and 2 electricity datasets. Validation of the mappings were conducted by relevant utility companies and feedback has been used to revise the mapping and utility thesauri (see Fu and Cohn (2008a) for more details).

3.3 Schema Integration

The two main problems in integrating heterogeneous utility data to a common data model are:

1. Defining the schema of the global model that is fit for purpose
2. Determining the mappings and transformations required to integrate source utility schemas to the global schema

Initially, automated and semi-automated techniques were employed to determine the global schema and its mappings. Unfortunately, the heterogeneities in the supplied data models meant that this approach was unsuccessful. Hence, the global schema was defined manually as described by Beck *et al.* (2008). This global schema, see Table 1, has been commented upon by utility partners and found to be robust for their purposes. Schema attributes have been placed into two categories: core and extended attributes. Core attributes are essential elements of the schema that are required by core end users (street workers, field engineers, back office planners).

Field Group	Global Schema Field	Short Definition	Data Type	Total mappings
Asset	serviceType	Service type: the type of service that the asset is carrying	Lookup	6
Asset	assetType	Asset Type: type of asset i.e. duct, pipe	Lookup	3
Asset	materialType	Material Type: what is the asset made from	Lookup	11
Asset	assetUseStatus	asset Use Status: in use, abandoned, not commissioned, planned	Lookup	12
Asset	assetSubType	Asset Sub Type: trunk main, distribution main	Lookup	12
Dimension	assetProfile	Asset Profile	Lookup	
Dimension	horizontalDiameter	Horizontal Diameter in mm	Double	
Dimension	verticalDiameter	Vertical Diameter in mm	Double	
Dimension	nominalDiameter	Nominal Diameter: expressed in the units of the underlying data store	text	14
Domain	assetDomain	Asset Domain: the utility domain the asset belongs to	Text	
Domain	assetOwner	Asset Owner: who owns the asset	Lookup	6
GIS	assetGisLink	original GIS Link	text	10
Location	assetTopBuriedDepth	Asset Buried Depth (to top of asset): below surface	Double	10

Table 1, The Core Global Schema

In order to populate the Global Schema with data, the relationship between fields in the source table and fields in the global schema require articulating. Many of these mappings are simple source-field to destination-field transpositions. However, a significant number of the mappings correspond to data transformations. These transformations can represent simple scaling of data, such as conforming to a pre-defined unit specification. However, more complex transformations require the use of multiple fields to generate an appropriate destination result. For example, in every utility domain except sewer the term ‘depth’ refers to the ‘depth of cover’ (a measurement from the surface to the top external measurement of the asset). In the sewer domain the term ‘depth’ refers to “depth to invert” (a measurement from the surface to the bottom internal measurement of the asset). To compare sewer depths with other utility depths one also requires information on diameter (or profile depth for non-circular assets) and thickness (if available). Integration is further complicated by the fact that the source data fields are, at times, sparsely and imperfectly populated. Therefore, on-the-fly data validation during the transformation process is required to ensure data quality. The mappings, transformations and validation components represent metadata that allow bespoke utility data models to interoperate at a schematic level via a mediating global schema.

The mappings and transformations are generated in conjunction with a domain expert from each utility company and held within the ‘Radius Studio’ software package developed by 1Spatial (www.1spatial.com). Radius Studio provides a toolset which allowed the rapid generation of complex data mappings and transforms. In future, we will be examining ways to convert this metadata into XSL Transformations. The ability to share this metadata through the Radius Studio web interface has allowed the VISTA team to rapidly validate and enrich the global schema in collaboration with our industrial partners.

Data is mapped or transformed using rules. A rule is a structured tree of hierarchical conditions, against which features can be tested. The rules are expressed in a form independent of any data store which means that rules can be easily re-used with different schema and data sources.

Rule formulation is best described with an example. Figure 2 is an artificial example used to attribute depth to a sewer pipe. The Global Schema field represents the average depth to the top of the asset. The source input polyline segment is 2d and has two attribute depths (upstream node and downstream node) and a diameter. The depth is measured in metres to the centre of the pipe and the diameter in mm. All source fields are sparsely populated. The example rule does the following:

- Checks if depth measurements have been populated.

- Calculates an average depth from the depth measurements and temporarily stores this value in AverageDepth.
- Divides the pipe diameter by 2, converts the units to metres and temporarily stores this value in PipeRadiusM.
- Populates the field GS_ASSETTOPBURIEDDEPTH with the value of ‘AverageDepth – PipeRadiusM.’

```

Description:
For
(
  SEWERS_POLYLINE objects: if
  SEWERS_POLYLINE.PIPE_DEPT0 equals 0 or
  SEWERS_POLYLINE.PIPE_DEPTH equals 0) then let AverageDepth =
  SEWERS_POLYLINE.PIPE_DEPT0 +
  SEWERS_POLYLINE.PIPE_DEPTH else {let SumOfDepth =
  SEWERS_POLYLINE.PIPE_DEPT0 +
  SEWERS_POLYLINE.PIPE_DEPTH and then let AverageDepth =
  SumOfDepth / 2} and then let PipeRadiusMM =
  SEWERS_POLYLINE.X_DIAMETER / 2 and then let PipeRadiusM =
  PipeRadiusMM / 1000 and then let
  SEWERS_POLYLINE.GS_ASSETTOPBURIEDDEPTH = AverageDepth -
  PipeRadiusM and then if (
  SEWERS_POLYLINE.PIPE_DEPT0 equals 0 and
  SEWERS_POLYLINE.PIPE_DEPTH equals 0) then let
  SEWERS_POLYLINE.GS_ASSETTOPBURIEDDEPTH = "No Depth
  Measurements"

```

Figure 2, Text view of a rule in Radius Studio (details have been removed to preserve anonymity)

3.4 Bridging the gap between semantic and schematic integration

The semantic mappings are incorporated into the integrated utility data by simply replacing the original utility values with the thesaurus terms. This produced a unified representation of the utility data (i.e. asset subtype, service type, and material type data) which is heterogeneous across the different utility domains. Some harmonisation was performed in order to overcome granularity mismatch. For example, for the asset type ‘water valve’ one utility may have a subtype ‘clockwise water check valve’ and another may have a subtype ‘water check valve’: the two codes are unified to water check valve. This may results in generalising a specific asset type to a more generic one, e.g. clockwise water check valve to water check valve. This preserves the correctness of the semantics. However, this approach is lossy, as we no longer know which item is the ‘clockwise water check valve’.

Radius Studio does allow the generation of validity rules in order to check the data conformance of the underlying data. These checks provide an overview of the underlying accuracy of a fully integrated dataset and insights into issues of spatial and attribute omissions and commissions. The Global Schema mapping has been undertaken and validated in collaboration with domain experts from each utility company. Hence, we consider that these mappings are valid at a company level for each sub-domain (gas, sewer, water etc.). However, we need to ensure that these mappings are still valid when conflated to the sub-domain and cross domain levels.

Validity at the sub-domain and cross domain levels means that all the integrated data from participating utility companies maintain semantic coherence and have a consistent degree of granularity. This will allow meaningful spatial and attribute queries to be articulated by the GIS or database without introducing errors of omission or commission from the Global Schema mapping process (or if these are introduced to ensure they are transparent to the end user). The ontology will be a significant tool for ensuring data validity at these levels.

3.5 The integration process

We have discussed the individual approaches for resolving each type of heterogeneity. Generic PL/SQL code was developed in Oracle to resolve the schematic and semantic heterogeneities. Figure 3 provides a high level view of the integration process.

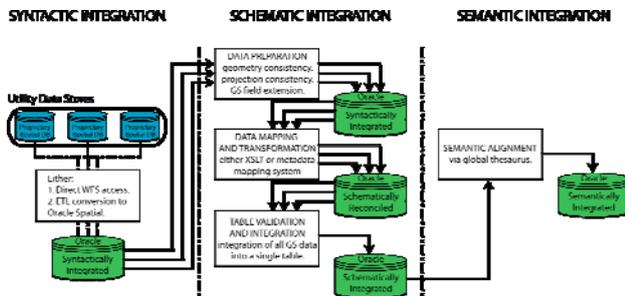


Figure 3, The proposed VISTA data integration flowline

The Oracle Spatial database is populated with source utility data. Each spatial table is checked for consistent geometry and the appropriate projection. Where problems are encountered these are automatically corrected. Each spatial table is extended with local copies of the Global Schema fields. These fields are then populated by invoking the Radius Studio web service to materialise the results of the metadata mappings. All the tables are then integrated into a single 3-d table. Any 2-d data is transformed into 3-d geometry on-the-fly. The semantic reconciliation procedures are then invoked.

3.6 Future developments for the integration framework

The work reported here represents the first steps towards generating an integrated framework for the delivery of utility data in the UK. It has ‘proved the concept’ that the dynamic integration of heterogeneous utility data sources is feasible. A cross-domain utility thesaurus has been developed and utilised to resolve semantic heterogeneities. Future research avenues will examine enriching the thesaurus and transforming it into an ontology. A robust cross domain ontology could have profound implications for the integration process itself. Instead of relying on a schema based integration process, an ontology driven integration process could be developed by mapping source data directly to the reference ontology using a bridge like D2RQ (Bizer and Cyganiak, 2007). This may offer further research potential for (semi)-automated matching and for visualizing semantic granularity (i.e. ontology based visualization). Once the ontology has been verified by the utility industry then the whole system will be modelled in UML. This may provide a framework for EU wide utility modelling under INSPIRE.

The ontology should allow the development of a lossless technique for resolving semantic heterogeneity. This would ultimately allow utility assets in different domains to be represented using the semantics and styling of any participating utility company. This moves closer to the interoperable framework discussed by Pundt (2008) employing the technologies described by van Harmelen (2008)

4. END-USER INFORMATION DELIVERY

The previous section has described the framework used to reconcile syntactic, semantic and schematic heterogeneities. The research programme also examined different techniques to deliver this integrated data to end-users. Two specific strands were pursued:

- Traditional webGIS delivery based upon deploying secure web services to established utility portals.
- Bespoke delivery using innovative end-user tailored visualisations.

There are also other delivery options. Two techniques that would require further evaluation are; direct access to WFS from a ‘thick client’ desktop CAD/GIS and an enhanced web-service to automatically process e-service requests (such as e-conveyancing: www1.landregistry.gov.uk/e-conveyancing). The former technique will provide back-office planners access to all the currently available utility data within their design and development platform of choice. This would allow conflicting utility location issues and risks to be evaluated and potentially mitigated during design. This has a range of obvious benefits.

4.1 Developing a utility web service

The VISTA consortium undertook 3 pilot projects. 2 using data from VISTA partners in the East Midlands (Lincolnshire, Nottinghamshire, Derbyshire, Leicestershire, Warwickshire, Northamptonshire and Rutland) and one, supported by the Scottish Executive, focussing on the Perth and Kinross region of Scotland (with data supplied by Scottish Water, Transport Scotland and Perth and Kinross Council).

A suite of Web Feature Services (WFSs) were deployed on a Leeds University server using the open source GeoServer package connecting to integrated utility data held in Oracle. Different materialized views were created in Oracle based upon data filters (utility data filtered by company or filtered by company *and* domain). Using this WFS infrastructure many different clients could consume integrated utility data from a single WFS environment.

A connection was established between the Leeds and client servers by IP filtering. The clients consume the WFS and render the output directly within their own WebGIS systems. All the pilot project systems have in-built authentication and security brokering. Although these pilots are on-going and have successfully implemented the technology, some problems were encountered.



Figure 4, Screen capture of a utility web service

Each consuming organisation used different software to access the WFS which resulted in connectivity and performance problems. Two organisations had problems connecting to the WFS. The former was because the service was password protected and the latter was due to strict XML parsing within the consuming software. These issues were subsequently resolved. The other organisation had no problem connecting to

the WFS (see Figure 4) but there were problems in performance which are still under investigation.

In summary although WFS is a standard, different implementations exist. In time, the resulting problems are likely to be resolved.

4.2 Bespoke utility visualizations

Utility companies rely on their own GIS/CAD systems to produce maps for on-site and back office work. Current maps do not address specific representation issues such as dealing with uncertain information and complex multi-domain graphics. Our visualization work moves away from the ‘one-map-fits-all’ solution by developing data, user and task specific mapping (for example to address different accuracy requirements and levels of detail (Boukhelifa and Duke, 2007; 2009)).

4.2.1 Uncertainty visualization

As we have already discussed, utility data has a range of associated spatial and attribute uncertainties. VISTA has focussed on representing uncertainties associated with positional accuracy. Information on positional accuracy is particularly important for asset detection because it allows more informed decisions in the field. We used two popular visual variables to indicate the positional accuracy of assets: blur and colour. The choice of depiction method was driven by the need for simple methods and visual metaphors suitable for non-technical audiences



Figure 5, Uncertainty visualization using blur and colour

Blurring (Figure 5 left) provides users with qualitative information about spatial accuracy. The more blurred a polyline, the less accurate its position. Our second scheme is the ‘‘traffic lights’’ visualization (Figure 5 right). It uses a three-colour unified scheme to paint colour bands around utility pipes indicating the confidence in the location of assets.

We encountered a number of issues when implementing the blur method; (a) perceptual issues related to the number of levels that the user is able to distinguish and remember; (b) over-plotting in 2D can result in a number of certain lines looking less blurred or more certain than they are in reality; (c) blurred lines on printed plans may be associated with low quality printing rather than low quality information. The ‘‘traffic light’’ metaphor is intuitive but mapping different categorizations of confidence from various data sources inevitably introduces inaccuracies, i.e. uncertainty about uncertainty.

4.2.2 Clutter and visual complexity

Multi-service plans are typically complex drawings due to close geographical proximity between assets, line crossings and busy junctions, missing 3D information and overlaps between labels. The problem of clutter is complicated by the lack of standardisation in the utility industry; guidelines for recording and displaying information on assets do not cover methods to

deal with detail. Sometimes, however, utility organizations adopt simplification or abstraction procedures manually to reduce clutter (e.g. Figure 6).

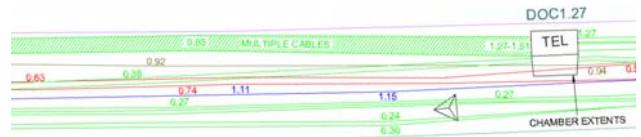


Figure 6, An example of simplification using hatching of a cluttered area of telecoms cables

Data integration should not lead to view confusion; our challenge is to find appropriate abstraction and simplification methods that improve visual understanding. We have considered clutter as an important factor in diminishing the aesthetics of the presented image. In our investigation to eliminate clutter in service plans, we took reductionist steps from graph drawing to measure clutter primitives such as proximity, bends, crosses and angles. We implemented proof-of-concept techniques to reduce clutter automatically by repackaging details and using aesthetics from information visualization. We argue that ‘de-cluttering’ does not mean loss of information, but rather repackaging details to make them more accessible.

5. PRACTICAL CONSIDERATIONS FOR IMPLEMENTATING A DATA INTEGRATION FRAMEWORK

The framework for integration described in the previous sections is an academic construct built in order to prove the concept that utility data integration is feasible. It was not designed to be exhaustive, nor was it expected to represent the range of delivery and integration solutions available. Integrating multiple datasets managed by different companies in different utility domains was expected to raise a number of practical problems. Recognition of these issues is considered important as they can frame the final architecture and may influence the approaches taken during implementation.

5.1 Data stability and operational impact

Previous experiments in integration have included either centralised database systems or organisationally independent systems based on the same data schema. Both of these options require a company to significantly change its data structure. Full harmonization of the underlying data models in this manner requires an unprecedented conceptual shift by participating utility companies which, given the costs involved and the fact that the models may not be suitable for their business processes, is highly unlikely. As stated by Lehto (2007) ‘It can even be argued that an organization should not change its internal data model on the basis of outside requirements, as these external demands are inherently diverse, possibly contradictory, and change over time’. Thus, these approaches, although theoretically sound, have failed in practice. A number of our utility partners initially understood integration to mean the above.

In order to reduce barriers to acceptance and participation it is important that any system has a low operational impact on working practices. In particular it should not require participants to change the format or structure of their underlying asset database holdings. Where changes are proposed then these should have additional business benefits.

For example, syntactic interoperability is seen as a potential barrier to successful implementation. A number of techniques can be used to achieve syntactic interoperability such as the use of ETL middleware. However, a more appropriate approach would be to access OGC compliant syntactically interoperable data in the form GML either directly or via WFS.

The use of WFS or GML has a number of benefits:

- The utility company retains full autonomy of its primary data store.
- The underlying data store can be changed with only minimal impact on the framework.
- Operational activity is not directly affected.
- Only the attributes required by MTU/VISTA will be exported, ensuring the security of non-essential, but potentially sensitive, data.

WFS and GML technologies are already being used by some utility companies to provide in-field GIS updates to their engineers. When implemented these field packages offer bi-directional transactional updates between the field device and the corporate asset repository.

5.2 Data currency

In order to make well-informed decisions based on utility data one should be aware of the currency of the data being analysed. Although preferable, it is not necessary that the data is up-to-date. Rather, the currency of the data should be appropriate for the application (i.e. fit for the end-users purpose). Data currency requirements vary with different user groups. For example, data currency is more of an issue for street workers, who will be engaging with the physical assets in the short term, than for back-office planners who may be planning works which will be installed months, or even years, in the future.

It should also be recognised that corporate GIS servers do not reflect the actual state of the asset network at any one point in time. There is always some degree of time lag between remedial work or the laying of new assets and the time when these appear on the corporate GIS. This delay can vary from less than 24 hours in a fully digital data collection system to weeks or months.

The appropriate appreciation of time lag means that different integration strategies can be developed. This is, in part, dependent on how the integration framework is implemented. If a virtual approach to implementation is taken then the integrated results would have the same currency as the source utility data. However, if a materialised approach is implemented, in order to deliver large scale integrated data, then data source specific refresh cycles can be devised that would significantly reduce processing overhead. This refresh cycle may also be driven by the type of 'role' a querying user possesses.

5.3 Ramifications for the integration architecture

Some of these issues can be contradictory at implementation. For example, if one was integrating data on-the-fly then accessing individual source datasets and disseminating nationwide utility data within an hour is impractical. The software architecture which is generated from VISTA will in part be structured by these practical considerations. Ultimately the utility industry will choose the appropriate architecture and,

unless there is a statutory requirement, each utility company will determine whether to make their data available through such a framework. However, in order to make this choice the industry needs to be made aware of the pros, cons, financial and organisational implications of any framework (see Table 2).

	Virtual	Materialised
Scale of Integration	Potentially Low	Full
Response Time	Delay	Real Time
Integration Overhead	Relatively Low	Very High: GRID
Data Currency	Current Snapshot	Almost Current Snapshot
Data Security	Data <i>not</i> Realised Locally	Data Realised Locally

Table 2, Comparison of Virtual and Material architectures

Two generic types of framework are postulated: Virtual and Materialised. The choice of either a Virtual or Materialised integration system has ramifications for the following issues:

1. Scale of integration: How much data can be sensibly integrated.
2. Response time: How long it takes for a user to get results.
3. Integration overhead: The processing resources required to integrate data.
4. Data currency.
5. Data security.

The strength of the virtual approach is that it reflects the data held in the source repositories at the time of querying (i.e. data currency) and that data need not be permanently stored (lower security risk). The weakness of this approach is that response time is relatively slow. The response time is dependent upon the quantity of data to be integrated, the data transfer speed and the data integration processing time. As a rule of thumb: the larger the spatial extent the longer the response time. Therefore, a virtual approach is potentially not appropriate where a significant quantity of data is required.

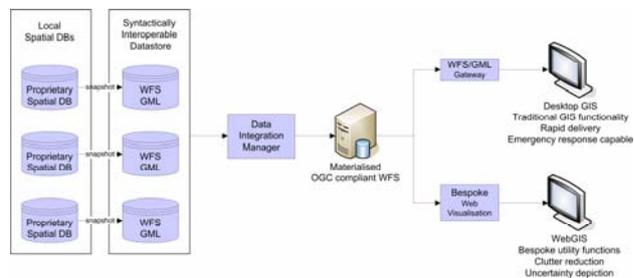


Figure 7, Materialised integration architecture

In the materialised approach data sources are merged into a single database, which is maintained centrally (see Figure 7). Queries are expressed against the integrated schema, but without directly accessing the local databases. The central database is periodically refreshed by globally updating all the source datasets at a periodic interval. The strength of the materialised approach is that it has very rapid response times and can deliver large datasets. The weaknesses of this approach are that the data is not always up-to-date, it is dependent upon the refresh cycle of the global update process, and all the source utility data will be permanently materialised on the server which may have security implications. Finally, in the materialised approach very large spatial datasets will be integrated. Efficiency gains might be realised by using GRID architecture.

There is, however, a third way which can combine the benefits of virtual and materialised architectures: GML "change only

updates". Instead of synchronising distributed data sources by re-integrating the full dataset one could integrate only the data that has changed (new data, updates and deletions) since the last update. This significantly reduces the quantity of data that is transferred. The OS use this form of delivery for updating Mastermap.

6. CONCLUSIONS

This paper has outlined the approaches used by the VISTA and MTU projects to resolve syntactic, schematic and semantic heterogeneity in the UK utility domain. We have also examined how this integrated data could be delivered to end-users and some of the visualization issues that structure how integrated utility data is provided to consumers. In order to enhance the probability of successful implementation we have considered a range of factors that may influence implementation. This does not directly impact on the design of the integration tools that are under development, but it will impact on how they are configured. The utility industry will need to decide what access it requires to competitors' data (size of data, response time, data currency etc.). With this information it is possible to implement an appropriate virtual or materialised delivery architecture.

7. ACKNOWLEDGEMENTS

We gratefully acknowledge the financial support of EPSRC (grant EP/C014707/1), the UK Technology Programme (project 15820) and the in kind contributions from VISTA partners. Thanks are due to the VISTA partners and the attendees of VISTA/MTU workshops for their insights into the business requirements of the different utility domains.

8. REFERENCES

- Aerts, K., Maesen, K. and Von Rompaey, A., 2006. A practical Example of Semantic Interoperability of Large-Scale Topographic Database using Semantic Web technologies, 9th AGILE International Conference on Geographic Information Science, Visegrád, Hungary.
- Ahmed, S., Kim, S. and Wallace, K.M., 2007. A Methodology for Creating Ontologies for Engineering Design. *Journal of Computer and Information Science in Engineering*, 7: 132-140.
- Beck, A.R., Fu, G., Cohn, A.G., Bennett, B. and Stell, J.G., 2008. A framework for utility data integration in the UK. In: V. Coors, M. Rumor, E.M. Fendel and S. Zlatanova (Editors), *Urban and Regional Data Management - Proceedings of the Urban Data Management Society Symposium 2007* Taylor & Francis, London, pp. 261-276.
- Bernard, L. et al., 2003. Ontologies for Intelligent Search and Semantic Translation in Spatial Data Infrastructures. *Photogrammetrie – Fernerkundung – Geoinformation*, 6: 451-162.
- Bishr, Y., 1998. Overcoming the Semantic and Other Barriers to GIS Interoperability. *International Journal of Geographical Information Science*, 12(4): 299-314.
- Bizer, C. and Cyganiak, R., 2007. D2RQ - Lessons Learned, <http://www.w3.org/2007/03/RdfRDB/papers/d2rq-positionpaper/>, 12th January 2009
- Boukhelifa, N. and Duke, D.J., 2007. The Uncertain Reality of Underground Assets, Visualization and Exploration of Geospatial Data, Stuttgart.
- Boukhelifa, N. and Duke, D.J., 2009. Visualizing Heterogeneous utility data: a case for aesthetic design, Eurographics Association, Munich.
- Cai, G., 2007. Contextualization of Geospatial Database Semantics for HumanGIS Interaction. *GeoInformatica*, 11: 217-237.
- Deliiska, B., 2007. Thesaurus and Domain Ontology of Geoinformatics. *Transactions in GIS*, 11(4): 637-651.
- Donaubauer, A., Straub, F. and Schilcher, M., 2007. mdWFS: A concept of Web-enabling Semantic Transformation, 10th AGILE International Conference on Geographic Information Science, Aalborg University, Denmark.
- Fu, G. and Cohn, A.G., 2008a. Semantic Integration for Mapping the Underworld, *Proceedings of GeoInformatics2008*, Gaizhuong, China.
- Fu, G. and Cohn, A.G., 2008b. Utility Ontology Development with Formal Concept Analysis, *Proceedings of the 5th International Conference on Formal Ontology in Information Systems*, Saarbrücken, Germany.
- Gruber, T.R., 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2): 199-221.
- Klien, E., Fitzner, D.i. and Maué, P., 2007. Baseline for Registering and annotating Geodata in a Semantic Web Service Framework, 10th AGILE International Conference on Geographic Information Science, Aalborg University, Denmark.
- Lehto, L., 2007. Schema Translations in a Web Service Based SDI, 10th AGILE International Conference on Geographic Information Science, Aalborg University, Denmark.
- Lemmens, R. et al., 2007. Enhancing Geo-Service Chaining through Deep Service Descriptions. *Transactions in GIS*, 11(6): 849-871.
- Marvin, S. and Slater, S., 1997. Urban infrastructure: the contemporary conflict between roads and utilities. *Progress in Planning*, 48(4): 247-318.
- NUAG, 2006. Capturing, recording, storing and sharing underground asset information – A review of current practices and future requirements. Report 06/WM/12/13, UKWIR.
- OS, 2007. Positional accuracy improvement programme, <http://www.ordnancesurvey.co.uk/oswebsite/pai/>, 30th May 2007
- Pinto, H.S. and Martins, J.P., 2004. Ontologies: How can they be built? *Knowledge and Information Systems*, 6(4): 441-464.
- Pundt, H., 2007. From Idea toward Ontology, 10th AGILE International Conference on Geographic Information Science, Aalborg University, Denmark.
- Pundt, H., 2008. The semantic mismatch as limiting factor for the use of geospatial information in disaster management and emergency response. In: S. Zlatanova and J. Li (Editors), *Geospatial information Technology for Emergency Response*. Taylor and Francis, London.
- Studer, R., Benjamins, V.R. and Fensel, D., 1998. Knowledge Engineering: Principles and methods. *Data & Knowledge Engineering*, 25: 161-197.
- van Harmelen, F., 2008. Semantic Web Technologies as the Foundation for the Information Infrastructure. In: P. Van Oosterom and S. Zlatanova (Editors), *Creating Spatial Information Infrastructures: Towards the Spatial Semantic Web*. CRC Press, London.