Published paper
Beck, A.R., Cohn, A.G., Bennett, B., Fu, G., Ramage, S., Sanderson, M., Stell, J.G. and Tagg, C (2008) *UK utility data integration: overcoming schematic heterogeneity.* In: Proceedings of SPIE, (7143). Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Geo-Simulation and Virtual GIS Environments., Art. No. 71431Z.

UK utility data integration: overcoming schematic heterogeneity

Anthony R. Beck*[a], Anthony G. Cohn[a], Michael Sanderson[b], Steve Ramage[b], Chris Tagg[b], Gaihua Fu[a], Brandon Bennett[a], John G. Stell[a]
[a]School of Computing, University of Leeds, Leeds, LS2 9JT, UK;
[b]1Spatial Ltd, Cavendish House, Cambridge Business Park, Cambridge, CB4 0WZ, UK

ABSTRACT

In this paper we discuss syntactic, semantic and schematic issues which inhibit the integration of utility data in the UK. We then focus on the techniques employed within the VISTA project to overcome schematic heterogeneity. A Global Schema based architecture is employed. Although automated approaches to Global Schema definition were attempted the heterogeneities of the sector were too great. A manual approach to Global Schema definition was employed. The techniques used to define and subsequently map source utility data models to this schema are discussed in detail. In order to ensure a coherent integrated model, sub and cross domain validation issues are then highlighted. Finally the proposed framework and data flow for schematic integration is introduced.

Keywords: Urban Infrastructure, Utility Assets, Heterogeneity, Data Integration, Schematic Reconciliation, GIS, Knowledge Exchange, Standards

1. INTRODUCTION AND RESEACH CONTEXT

Information exchange and re-use is crucial to many organisations throughout the world in order to reduce costs associated with data collection, assemblage and management. Privatised utility companies in the UK are one such example. The gas, water, sewer, electricity and telecommunications infrastructure is essential for modern urbanism and forms part of the fabric of both rural and urban landscapes. This apparatus is commonly located underneath, or is incorporated into, the built environment. Utility companies require quick and efficient access to their assets in order to assess or repair damage and replace any degraded material. Particularly in urban environments, the bulk of utility infrastructure is placed in the highway (road or pavement). As a consequence of the proliferation of services over time, the space beneath the highway is often congested with pipes, ducts, cables and other utility assets. This congestion causes serious problems when repairs are undertaken or new facilities are added to a network. This can result in third party damage and injury to the workforce.

Every year, in excess of four million holes are dug in UK roads to repair assets, provide connecting services to new premises and to lay new cables and pipes. Although recently installed assets may have been well mapped, location and attribute data on older services can be very poor, in some cases even non-existent. This poor data quality can lead to unnecessary holes dug in the wrong place and third party damage to other underground services. Equally important, there are also considerable indirect costs to society owing to disruption on the roads caused by works, waste and pollution[1]. These costs are born by individuals, businesses and government agencies. The core of the problem is that there is at present insufficient and inadequate knowledge about what is where. Existing information is not always used to its maximum benefit.

Although most utility companies maintain their asset data in digital vector formats[2, 3], it is common for third party users to only have access to printed versions (some of this data may be hardcopy derived from a vector web GIS). Furthermore, there can be a time-lag of many weeks between data request and receipt which can result in emergency excavations proceeding with very little, or outdated, collateral information. Therefore, current approaches to information sharing significantly reduce the richness of the source data and its fitness for a variety of end purposes.

It is postulated 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 street works in the UK by improving both the co-ordination of works and the quality of information which is shared. 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. The integration techniques proposed in this paper will not resolve incorrect or missing data. However, the integrated dataset can be used to model and infer instances of incorrect or missing data.

The School of Computing at the University of Leeds is researching into these issues in both the Mapping The Underworld (MTU: www.mappingtheunderworld.ac.uk) and Visualizing integrated information on buried assets to reduce streetworks (VISTA: www.vistadtiproject.org) projects. The VISTA project has 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. Amongst other things, these partners provide real world data for the project, as well as a wealth of experience in their respective fields. At present the VISTA project partners provide access to 16 datasets unique to company and domain. These datasets are used extensively in the modelling and design phases discussed in this paper.

VISTA is a proof of concept project that is evaluating the range of technical challenges and approaches to integrate utility data that exhibit syntactic, schematic and semantic heterogeneity. This paper focuses on the techniques used to resolve the schematic heterogeneity that exists in the UK utility domain in order to integrate disparate geospatial data. It is recognised that each utility company has their own approach to structuring their database schemas.

VISTA could resolve the integration problem by simply displaying the different utility data as individual layers in an online GIS or as multiple Web Feature Services (WFS). However, fully integrating the spatial and attribute data from all the utilities allows the production of more sophisticated representations which can make the information easier to understand for users and allows the development of new tools to better understand the provenance and accuracy of the information.

Previous integration approaches have tried to harmonise the source data schemas. However, utility companies have been naturally reticent to change their underlying data architecture[4, 5]. The work reported here explores the alternative approach of using a Global Schema based system[6, 7]. This Global Schema must provide a framework to unify the utility data from each of the partner domains (electricity, gas, sewer, telecoms and water), without requiring a change in the source schema, while providing enough flexibility to articulate the visualization and analytical requirements of each domain. Automated and semi-automated approaches to determining the Global Schema were found to be inappropriate since the data structures were too heterogeneous. Therefore, a manual approach was used. In order to translate data in the source databases to the Global Schema mapping metadata needs defining. This complex metadata is maintained within Radius Studio software developed by 1Spatial. The mapping metadata is used to populate Global Schema fields extended onto the source tables. A generic approach to validating the spatial conformity and appending these spatial databases into a destination table has been developed.

This paper will discuss the range of heterogeneities in the utility domain and then focus upon how the schematic differences can be reconciled using a Global Schema based architecture. Although, automated approaches to Global Schema definition were attempted the heterogeneities of the sector were too great: hence we will describe how we defined the Global Schema using a manual approach. The techniques used to define and subsequently map source utility data models to this schema are discussed in detail. In order to ensure a coherent integrated model, sub and cross domain validation issues are then highlighted. Finally the proposed framework and data flow for schematic integration is introduced.

2. HETEROGENEITIES IN THE UTILITY DOMAIN

For utility assets held in a digital format, the differences in data systems, structures and formats limits the ability to integrate data from different utilities effectively. This has the potential to hinder its usefulness in street works and has been recognised by the National Underground Assets Group[8]. The heterogeneities are caused by many factors but the main reason is that utility knowledge and data is typically autonomous, i.e. created and maintained independently by individual utility companies. Furthermore, the data is encoded in an uncoordinated way, i.e. without consideration of compatibility and interoperability with other utility systems. This practice is understandable as the principal remit for digitising assets is to improve operational systems for the company and not to improve data sharing within the sector[4]. This means that different companies have different abstracted views of reality which consequently impacts on asset data models and recording systems. Overcoming these heterogeneities is an essential first step to achieve utility integration and move towards domain interoperability. This will provide a foundation for sharing across the spatial data supply chain or within any spatial data infrastructure.

According to our investigations and reports from UK utility companies, this group of heterogeneities covers a wide range of issues, from the data models, to the underlying data and information that are being stored. For the purpose of discussion, we classify heterogeneities associated with utility records into the three categories discussed by Bishr[9]: syntactic, schematic and semantic heterogeneity.

2.1 Syntactic Heterogeneity

Syntactic heterogeneity refers to the difference in data format[10]. The same logical model can be represented in a range of different physical models (for example ESRI shape file or Geography Mark-up Language (GML)). The treatment of spatial data varies greatly, from compressed binary data (such as a scan), to data models specifically designed for spatial data[11]. This mismatch between underlying data models implies that the same information could be represented differently in different utility systems. The most profound difference is in the storage paradigm: relational, object orientated and hybrids.

Partner utility companies rely on a range of different GIS including GE Smallworld, ESRI ArcMap, AutoDesk Map and MapInfo, employing a range of storage solutions including Oracle, SQL Server and ArcSDE. A mechanism to resolve this syntactic heterogeneity has yet to be devised but is likely to employ either Open Geospatial Consortium (OGC) approved formats or potentially lossy third party interoperability solutions (Extract, Transform and Load (ETL) tools) such as Feature Manipulation Engine (FME). However, some utility companies are starting to make their data available in OGC approved syntactically interoperable formats and services such as GML and Web Feature Service (WFS). At present utility companies have configured these systems to aid field engineers; as these users demand high quality geometry and a rich data model it is likely that such frameworks will be appropriate for integration purposes.

Syntactically interoperable formats underpin a number of geospatial integration frameworks currently under development[12]. These are predominantly based on Service Oriented Architectures (SOA)[13-15]. For the purposes of this paper it is assumed that all source data is available in a syntactically interoperable format.

The proposed global schema approach is articulated within an Oracle Database 10g architecture. This architecture was chosen as Oracle Spatial is Open Geospatial Consortium (OGC) compliant, is highly scalable and is grid enabled, which may be essential for processing and integrating very large nationwide data. Each utility dataset was imported into Oracle Spatial using Safe Software's FME software.

2.2 Semantic Heterogeneity

Semantic heterogeneity refers to differences in naming conventions and conceptual groupings in different companies[9]. Data, or instance, level semantic heterogeneity can be subdivided into naming and cognitive heterogeneities. Naming (synonym) mismatch arises when semantically identical data items are named differently[16, 17]. 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 word objects from different perspectives. Cognitive (homonym) mismatch arises when semantically different data items are named identically in different utility systems. 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[10]. Reconciling these cognitive heterogeneities is more difficult but is achievable through ontology mapping.

The degree of schematic granularity can further affect the results. For example two different companies lay exactly the same water pipe, one company refers to the material at a fine level of granularity (MDPE – Medium Density PolyEthylene) and another refers to the material at a coarse degree of granularity (Plastic). Both terms are correct and fit for the internal business purposes but do present a challenge for semantic integration.

The semantic component of the VISTA project is discussed by Fu and Cohn[18].

2.3 Schematic Heterogeneity

Schematic heterogeneity refers to the differences in data model between organisations. The database schema is designed at the conceptual modeling 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.

Heterogeneities can arise at this level in many forms due to the different domain perceptions, business logic and interests of different user groups. These all impact on the type of information recorded, the ways that this information is represented, the ways that different types of information relate to each other and various semantics attached to utility records. Common heterogeneity issues are detailed below:

- Structures: different utility databases have different record structures.
- Structural semantics: elements encoded at the schema level are usually attached with some data semantics. The following are some typical semantic heterogeneities existing among utility records:
  - type mismatch occurs when same class of data are assigned with different data types, e.g. one utility system may use a text field to record material type whilst another uses a numeric field.
  - range mismatch arises when different utility systems allow their data items to have different value ranges.
- Granularity: different systems encoding data at different levels of detail, e.g. one utility system encodes mains pipes whilst another also encodes service pipes.

What follows is a discussion on how the VISTA project is proposing to resolve schematic heterogeneity for the UK utility domain.

3. VISTA SCHEMATIC DATA INTEGRATION

A critical step in the integration process is to produce a single data model that enables asset data from multiple information sources to be represented in a common format. The two main problems in designing a common data model are determining the mappings between elements of individual utility data models (or database schemas), and integrating them into a unified model based upon these mappings.

A number of tools are available that automate or semi-automate schema integration, producing a global schema and the mappings from the source data models to the resultant global schema. It is recognised that it may not be possible to determine all mappings between schemas automatically, primarily due to the fact that the matching clues are often unreliable and incomplete, and do not provide sufficient information to determine the exact nature of the relationship. However, the manual creation of mappings is laborious and error-prone, and therefore many studies have investigated faster schema matching techniques.

Two integration tools were chosen to aid in the determination of the global schema: PROMPT[19], an extension of the ontology editor PROTÉGÉ, and COMA+[20]. These two tools were considered representative of the range of available tools. The former employs a semi-automated approach with a simple linguistic matching algorithm. The latter is a more complex automated system employing richer matching criteria and composite searches. The success of automated schema integration is largely determined by the level of heterogeneity of the schemas to be integrated. The greater the schema heterogeneities the lower the chances of successful automated integration. Unfortunately, both PROMPT and COMA+ largely failed to automatically produce meaning integration metadata using trial utility data[21]. The main reason for this is the high level heterogeneities of these data models. For example:

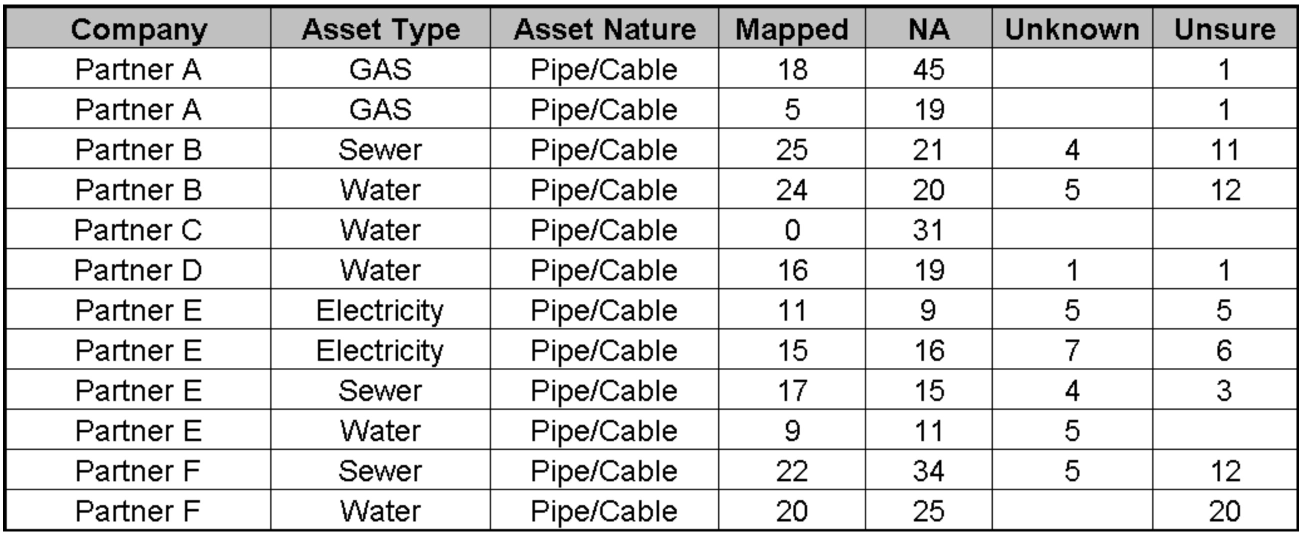
- Companies use different conventions to name the same feature classes.
- Companies encode different attributes for the same feature class.
- Lookup tables have, paradoxically, reduced the semantic richness of some of the data models.

To resolve this, an obvious solution is to develop a better integration tool. However, COMA+ is one of most powerful tools available. Therefore designing a more powerful tool may still not be able to solve the problems. Hence, integration will be performed manually or at best semi-automatically. A manual approach was taken as this provided an opportunity for the project team to acquire deep domain knowledge. This knowledge may prove essential if the tool is realized in the market place.

3.1 Generating the Global Schema Manually

Each of the original databases had a range of different asset records for each domain. Although each of these record types was nominally structured by their geometry (mainly polyline for pipes, points for street furniture), the differences in representation between the utility companies was significant. Some companies held a single point, polyline and polygon spatial database and relied on the attributes to distinguish between the different asset types. Other companies provided multiple spatial databases corresponding to the different asset types in their network (each with their own set of attribute information). For practical reasons the domain of the problem was reduced by focusing only on principal pipe/cable datasets from each of the utility companies (see Table 1).

Table 1 Data used for Global Schema matching



Company	Asset Type	Asset Nature	Mapped	NA	Unknown	Unsure
Partner A	GAS	Pipe/Cable	18	45		1
Partner A	GAS	Pipe/Cable	5	19		1
Partner B	Sewer	Pipe/Cable	25	21	4	11
Partner B	Water	Pipe/Cable	24	20	5	12
Partner C	Water	Pipe/Cable	0	31		
Partner D	Water	Pipe/Cable	16	19	1	1
Partner E	Electricity	Pipe/Cable	11	9	5	5
Partner E	Electricity	Pipe/Cable	15	16	7	6
Partner E	Sewer	Pipe/Cable	17	15	4	3
Partner E	Water	Pipe/Cable	9	11	5	
Partner F	Sewer	Pipe/Cable	22	34	5	12
Partner F	Water	Pipe/Cable	20	25		20

A database was created that summarised the nature of each asset type and recorded the field names, data types and value examples for each field in the supplied physical model of the spatial databases. Using the supplied metadata (logical model and other supporting documentation) logical mappings and explicit definitions were added to these records wherever possible. A key issue in resolving heterogeneity is the acquisition of appropriate metadata and discerning the semantic relationships between constructs of the different database schemas. Variable levels of metadata were provided by the utility companies: when limited or poor metadata was available then matching was made more difficult.

After evaluating the information from each of the different utility domains (with the exception of telecoms) a preliminary Global Schema was designed. This schema selected fields that were considered important for street works and back office planning and used the recommendations from NUAG and UKWIR[8, 22] (see Table 2). Significant time was spent determining names for the fields in the Global Schema. It is essential that the field names are semantically transparent and retain meaning throughout each of the utility sub-domains. The individual fields from the physical models were then coarsely mapped to the global schema. These coarse mappings did not represent the final mappings for data integration. Rather, these mappings indicated relationships between the source and Global Schemas which could be used to evaluate the nature of the mappings. A Global Schema field was mapped to each field in each source table. Where a field was not considered important it was given the value 'NA'. Those fields with ambiguous definitions or metadata that may be important were given the value 'unsure'. Where there was not enough information to accurately map the field it was given 'unknown' (this information is summarised in Table 1). In some instances many fields in the utility database were mapped onto one field in the Global Schema. Data from Partner C has not yet been mapped owing to difficulties in interpreting the fields in the physical model (this company is in the process of redefining and migrating its asset management GIS).

Transformation issues were recorded for each field. Two principal types of transformation issue were encountered:

- Consistency reconciliation: how units or measurements require transforming for a consistent representation. For example, all depths/height should be to the top of the asset.
- Data Dictionary (lookup table) reconciliation: how different data dictionaries can be merged to generate a global utility domain data dictionary.

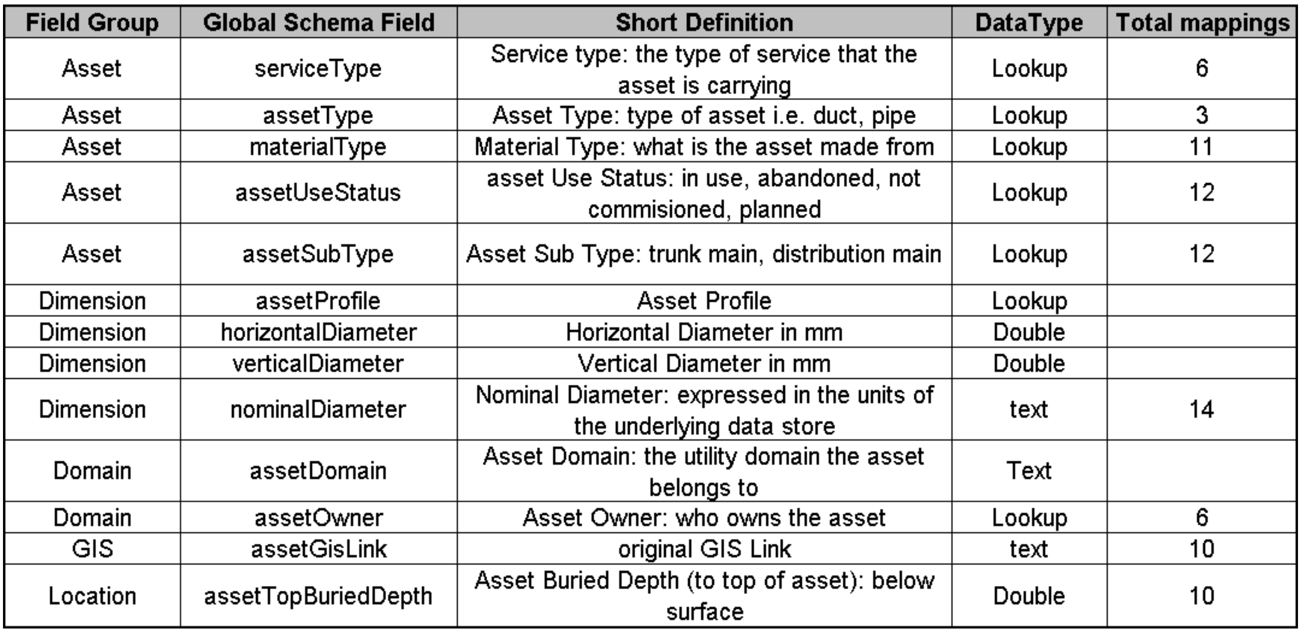
Consistency reconciliation provided two specific challenges: 3d measurements and conversion of imperial measurements to metric units. Although the majority of GIS records are 2d in nature, the third dimension is commonly articulated as depth or elevation attributes. For the majority of domains depth and elevation refers to ‘Depth/Elevation of Cover’ (the depth or elevation of the top of the asset). However, as gravity plays an important role in transportation for the Sewer domain depth and elevation is commonly cited as ‘Depth/Elevation of Invert’ (the lowest point of the sewer channel). The invert point is an internal pipe measurement whereas the cover point is an external pipe measurement. Therefore, in order to make sewer measurements consistent with all other domains one needs to convert an invert measurement to a cover measurement requiring information on vertical diameter and pipe thickness.

Imperial and metric measurements are both used to record pipes, ducts and other utility fittings. Reconciling imperial and metric measurements is more than just a matter of converting units, since each measurement affects other properties of the asset. For example a 3” pipe takes a different fitting to a 76.2mm pipe. This is of significant value to an on-site engineer. Therefore, the Global Schema contains a nominal diameter measurement field which expresses any measurement in terms of the units of the underlying source data and other fields which reconcile the measurements to metric units.

This global schema has been distributed to the utility partners for comment and feedback and found to be robust for the envisaged end-uses. Any of the ambiguities arising from ‘unsure’ or ‘unknown’ mappings have been resolved. The Global Schema has been continually revised: in the current version 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, etc.). Extended attributes enrich the data model but are not essential for its successful implementation. In fact some of the extended attributes, such as ‘dateAssetLaid’, may inhibit implementation as the utility company may not want to share such information. For simplicity and clarity, the rest of this paper refers to the core schema.

The core schema has been successfully tested to integrate network *and* furniture data from most of the utility domains. The success with furniture data is attributed to the generic design.

Table 2 Core Global Schema



Field Group	Global Schema Field	Short Definition	DataType	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 commisioned, 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

3.2 Mapping source data to the Global Schema

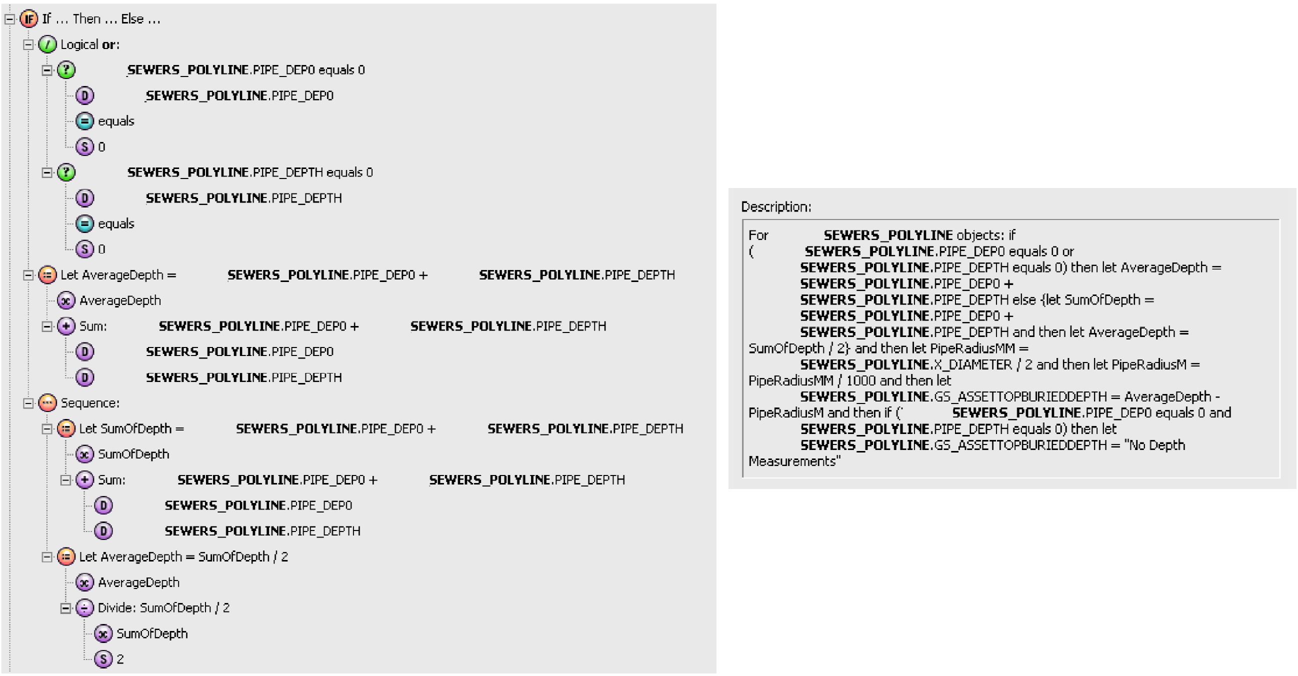
In order to populate the Global Schema with data, the relationship between fields in the source table and fields in the global schema require articulating. Many of these mappings are simple source-field to destination-field transpositions. However, a significant number of the mappings represent data transformations. These transformations can represent simple scaling of data, such as conforming to a pre-defined unit specification. However, some more complex transformations require the use of multiple fields to generate an appropriate destination result. 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 is an enterprise spatial data integration tool that simplifies the generation and management of complex transformation metadata. It also enables knowledge management through a unified rules repository. Radius Studio provides a toolset which not only allowed the rapid generation of complex data mappings and transforms but allowed the VISTA team to address data quality issues by creating a number of data validation rules. The ability to share this metadata through the Radius Studio secure 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 1 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



```
If ... Then ... Else ...
  Logical or:
    SEWERS_POLYLINE.PIPE_DEP0 equals 0
      SEWERS_POLYLINE.PIPE_DEP0
      equals
      0
    SEWERS_POLYLINE.PIPE_DEPTH equals 0
      SEWERS_POLYLINE.PIPE_DEPTH
      equals
      0
  Let AverageDepth = SEWERS_POLYLINE.PIPE_DEP0 + SEWERS_POLYLINE.PIPE_DEPTH
    AverageDepth
    Sum: SEWERS_POLYLINE.PIPE_DEP0 + SEWERS_POLYLINE.PIPE_DEPTH
      SEWERS_POLYLINE.PIPE_DEP0
      SEWERS_POLYLINE.PIPE_DEPTH
  Sequence:
    Let SumOfDepth = SEWERS_POLYLINE.PIPE_DEP0 + SEWERS_POLYLINE.PIPE_DEPTH
      SumOfDepth
      Sum: SEWERS_POLYLINE.PIPE_DEP0 + SEWERS_POLYLINE.PIPE_DEPTH
        SEWERS_POLYLINE.PIPE_DEP0
        SEWERS_POLYLINE.PIPE_DEPTH
    Let AverageDepth = SumOfDepth / 2
      AverageDepth
      Divide: SumOfDepth / 2
        SumOfDepth
        2
```

Description:

```
For SEWERS_POLYLINE objects: if
( SEWERS_POLYLINE.PIPE_DEP0 equals 0 or
SEWERS_POLYLINE.PIPE_DEPTH equals 0) then let AverageDepth =
SEWERS_POLYLINE.PIPE_DEP0 +
SEWERS_POLYLINE.PIPE_DEPTH else {let SumOfDepth =
SEWERS_POLYLINE.PIPE_DEP0 +
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_DEP0 equals 0 and
SEWERS_POLYLINE.PIPE_DEPTH equals 0) then let
SEWERS_POLYLINE.GS_ASSETTOPBURIEDDEPTH = "No Depth
Measurements"
```

Logical View

Textual View

Figure 1 Comparing the Logical and Textual views of a rule in Radius Studio (details have been removed to preserve anonymity)

Radius Studio has proven to be a robust platform for developing, managing and sharing the complex mapping and transformation metadata. Trying to develop this metadata directly as text, SQL statements or Extensible Stylesheet Language Transformations (XSLT) for XML data would have been unnecessarily unwieldy and would have been difficult to share coherently with domain experts.

3.3 Mapping validation

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 rendered 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 VISTA domain ontology, currently under development[18], will be a significant tool for ensuring data validity at these levels.

4. SCHEMATIC INTEGRATION FRAMEWORK AND DATA FLOW

The VISTA project team has developed a conceptual framework to support utility data integration. Beck *et al*[23] discuss this framework in detail and the pros and cons of a virtual or materialized implementation path. Irrespective of what approach is employed data integration will include a number of specific processing and validation steps. The processing flowline, focusing on schematic integration, is shown in Figure 2.

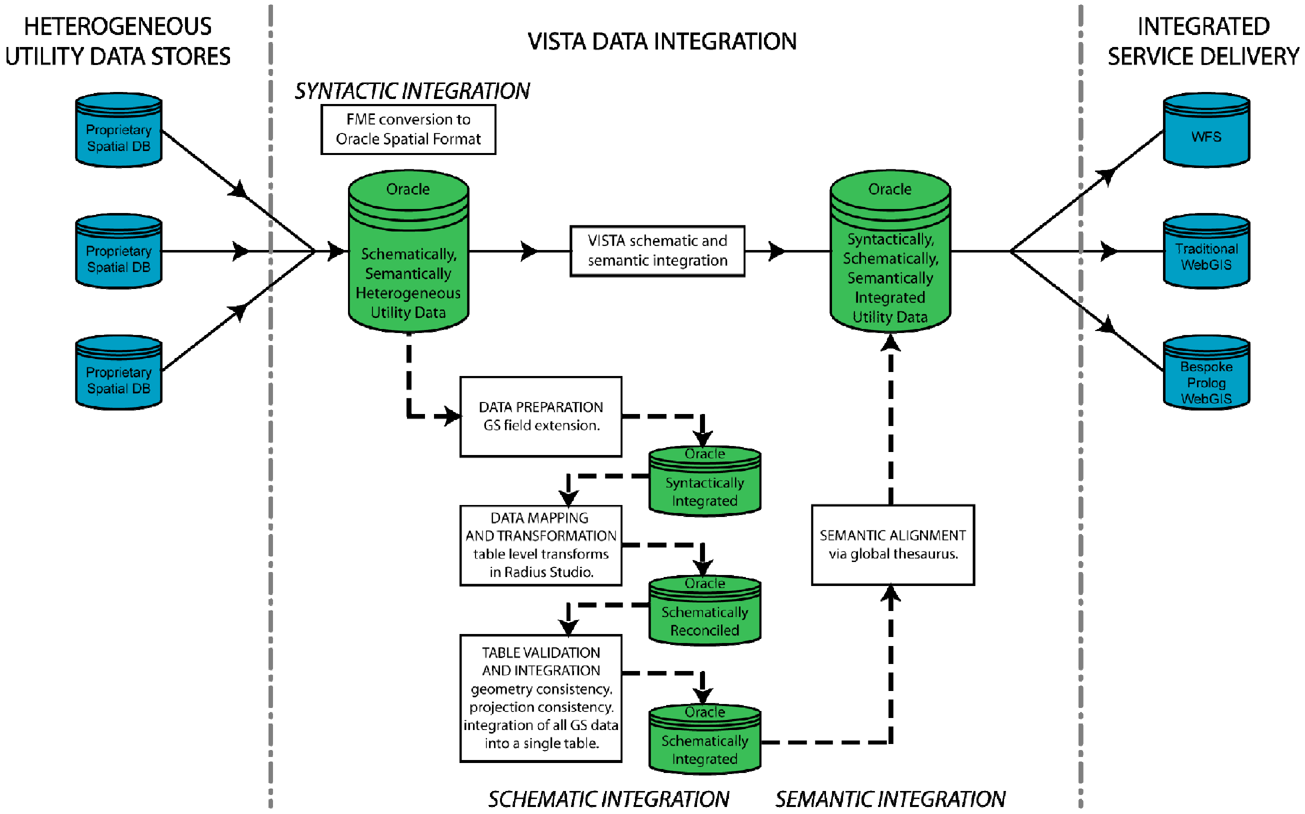

HETEROGENEOUS UTILITY DATA STORES
VISTA DATA INTEGRATION
INTEGRATED SERVICE DELIVERY
SYNTACTIC INTEGRATION
FME conversion to Oracle Spatial Format
Proprietary Spatial DB
Proprietary Spatial DB
Proprietary Spatial DB
Oracle
Schematically, Semantically Heterogeneous Utility Data
VISTA schematic and semantic integration
Oracle
Syntactically, Schematically, Semantically Integrated Utility Data
WFS
Traditional WebGIS
Bespoke Prolog WebGIS
DATA PREPARATION GS field extension.
Oracle
Syntactically Integrated
DATA MAPPING AND TRANSFORMATION table level transforms in Radius Studio.
Oracle
Schematically Reconciled
TABLE VALIDATION AND INTEGRATION geometry consistency, projection consistency, integration of all GS data into a single table.
Oracle
Schematically Integrated
SEMANTIC ALIGNMENT via global thesaurus.
SCHEMATIC INTEGRATION
SEMANTIC INTEGRATION

Figure 2 Data integration flowline demonstrating syntactic, schematic and semantic reconciliation in the proof-of-concept framework.

Schematic heterogeneity is reconciled using a number of steps:

- Data Preparation
- Data Mapping and Transformation in Radius Studio
- Table Validation and Integration

Initially each of the source utility data tables from the Oracle Spatial database is prepared for integration by extending the schema to contain a local version of the Global Schema fields. We have found it easier to validate the data transformed by Radius Studio if it can be compared directly to the source data at the record level prior to integration into a single data table.

The second stage is to populate the local Global Schema fields using the mapping and transformations developed in Radius Studio. The final stage integrates all the source utility datasets into a single table. However, in order to successfully update a target table in Oracle Spatial, all source tables must share the same dimensionality and spatial projection. The project requires three dimensional data projected to the British National Grid (Spatial Reference ID (SRID) = 27700). Metadata for all the source utility data tables is generated containing information on the Oracle schema and owner, the spatial field containing the geometry, metadata and table SRID and the table dimensionality. A PL/SQL procedure cycles through every record in the metadata table and applies projection corrections to each invalid SRID and extends the dimensionality of any non 3d dataset. Once the projection and dimensionality have been validated an 'insert into' SQL statement appends each source table into a single schematically integrated spatial table. The whole schematic reconciliation process is generic and has been automated using PL/SQL code (which makes use of the Radius Studio web services).

5. CONCLUSIONS AND FUTURE WORK

This paper has described the conceptual approach and practical advances to the schematic integration of utility data on the VISTA project. Once the mappings between the source data schemas and the global schema have been completed they will require binding to the thesauri/ontology being developed in Fu and Cohn[18]. The resultant mappings and semantic bindings will require validating. The process will not only require company and domain (intra-schema) validation by domain experts but also that the mappings are coherent when considered at the sub-domain and cross utility domain levels. The major issues in this phase will involve reconciling varying attribute granularities and conceptual relationships between sub-domains. Once all the structures, mappings and values have been validated then all the component processes required to successfully integrate heterogeneous utility data will have been completed. The final task is to embed these processes within a software architecture.

During the proof-of-concept phase, logical level mapping has been undertaken. However, if this project moves to an implementation phase it may be appropriate to define schema mapping at the conceptual level. Donaubauer *et al*[12, 13] discuss this approach using OGC compliant web services.

Inevitably, any VISTA solution will be web based and Service Oriented Architecture provides an appropriate solution for geospatial data. The challenge in the utility sector is to leverage business critical geospatial data to external consumers[5]. It is unlikely that any organization will change its internal data structures and business process to facilitate such integration. A solution is to schematically and semantically integrate utility at the service level. In this scenario, views of the source data are made securely available as web services, thus maintaining the primacy and integrity of the source data. Multiple utility services can then be virtually integrated on-the-fly or materialized by other accredited external web services.

It may be germane to ask why the VISTA project may succeed where other approaches have failed. We feel that previous approaches to utility data integration have either failed, or have not moved through to industry applications, for a combination of the following reasons that do not solely rely on cost:

- Technical issues
  - Lack of a robust high-speed communication infrastructure
  - Many asset records were still in a paper format
  - Immature hardware, software structures

    - Poor interoperability solutions
    - Lack of appropriate integration tools
- Organisational issues
    - Low level of industry commitment
    - Perceived security issues
    - The lack of a suitable catalyst to the drive integration processes

Over the past two decades most utility companies have dealt with converting their paper or raster material into a vector format and have enriched their data structures so that a variety of new information can be realised. This has been a common stage for all industries that manage geo-spatial data. The current challenge for the industry is to find ways to integrate these heterogeneous data sources as part of an emerging discipline referred to as Knowledge Management[24]. In response, interoperable file formats have been developed that allow information to be shared without data loss and tools are now available that can seamlessly integrate these data. However, as important as technical solutions are pull factors:

- The environmental impact of traffic congestion.
- The significant costs involved in rehabilitating infrastructure (some of which may be two hundred years old).
- Increased expectation from the public, industry, regulators and government that effective knowledge sharing occurs.

We are fortunate to be conducting this research at a time when many of the technical hurdles have been resolved: most utility companies store their asset information digitally and interoperable solutions, integration tools and high speed connective infrastructure are all available. Hence, mechanisms to integrate utility data are technically feasible at a relatively modest cost. Critically, legislation in the form of the Traffic Management Act is acting as a catalyst to gain industry wide commitment to initiatives which will facilitate data exchange. This is reflected in the breadth of industrial partners on the VISTA project: probably making it the most representative utility data integration project to date in the UK. Positive results from this project will provide a firm foundation for UK, and potentially international utility integration programmes. This project exemplifies the collaborative nature of solution building for cross-sector spatial information management.

ACKNOWLEDGEMENTS

We gratefully acknowledge the financial support of EPSRC under grant EP/C014707/1, the UK Department of Trade and Industry (Technology Programme project 15820) and the in kind contributions from VISTA partners (VISTA is a consortium consisting of over 20 partners (listed at www.vistadtiproject.org)). 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.

REFERENCES

1. McMahon, W., Evans, M., Burtwell, M.H., Parker, J.: The real costs of street works to the utility industry and society. UKWIR (2005)
2. Beck, A.R., Boukhelifa, N., Fu, G., Hickinbotham, S.J., Parker, J., Bennett, B., Cohn, A.G., Duke, D., Stell, J.G.: Utility data integration and knowledge representation in the UK: the VISTA project. In: Anand, S., Ware, M., Jackson, M., Vairavamoorthy, K., Abrahart, R.J. (eds.): Geohydroinformatics - Integrating GIS and Water Engineering. Taylor & Francis, London (in press)
3. Beck, A.R., Fu, G., Cohn, A.G., Bennett, B., Stell, J.G.: A framework for utility data integration in the UK. In: Coors, V., Rumor, M., Fendel, E.M., Zlatanova, S. (eds.): Urban and Regional Data Management - Proceedings of the Urban Data Management Society Symposium 2007 Taylor & Francis, London (2008) 261-276

4. Bernard, L., Einspanier, U., Haubrock, S., Hübner, S., Kuhn, W., Lessing, R., Lutz, M., Visser, U., "Ontologies for Intelligent Search and Semantic Translation in Spatial Data Infrastructures." Photogrammetrie – Fernerkundung – Geoinformation **6,** 451-162 (2003)

5. Lehto, L.: Schema Translations in a Web Service Based SDI. 10th AGILE International Conference on Geographic Information Science, Aalborg University, Denmark (2007)

6. Motro, A., "Superviews: Virtual Integration Of Multiple Databases." IEEE Transaction on Software Engineering **13,** 785-798 (1987)

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

8. NUAG: Capturing, recording, storing and sharing underground asset informa-tion – A review of current practices and future requirements. UKWIR (2006)

9. Bishr, Y., "Overcoming the Semantic and Other Barriers to GIS Interoperability." International Journal of Geographical information Science **12,** 299-314 (1998)

10. Aerts, K., Maesen, K., Von Rompaey, A.: 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 (2006)

11. Rigaux, P., Scholl, M., Voisard, A.: *Spatial Databases: With Application to GIS*. Morgan Kaufmann, 2001

12. Donaubauer, A., Straub, F., Schilcher, M.: mdWFS: A concept of Web-enabling Semantic Transformation. 10th AGILE International Conference on Geographic Information Science, Aalborg University, Denmark (2007)

13. Donaubauer, A., Fichtinger, A., Schilcher, M., Straub, F.: Model Driven Approach for Accessing Distributed Spatial Data Using Web Services - Demonstrated for Cross-Border GIS Applications. XXIII International FIG Congress, 8-13 Oct, 2006,, Munich, Germany (2006)

14. Klien, E., Fitzner, D.i., Maué, P.: Baseline for Registering and annotating Geodata in a Semantic Web Service Framework. 10th AGILE International Conference on Geographic Information Science, Aalborg University, Denmark (2007)

15. Lemmens, R., de By, R., Gould, M., Wytzisk, A., Granell, C., van Oosterom, P., "Enhancing Geo-Service Chaining through Deep Service Descriptions." Transactions in GIS **11,** 849-871 (2007)

16. Klien, E., Lutz, M., Kuhn, W., "Ontology-based discovery of geographic information services - An application in disaster management." Computers Environment and Urban Systems **30,** 102-123 (2006)

17. Bernard, L., Einspanier, U., Haubrock, S., Hübner, S., Klien, E., Kuhn, W., Lessing, R., Lutz, M., Visser, U.: Ontology-Based Discovery and Retrieval of Geographic Information in Spatial Data Infrastructures. Geotechnologien Science Report, Vol. 4. Koordinierungsbüro Geotechnologien, Potsdam (2004) 15-29

18. Fu, G., Cohn, A.G.: Semantic Integration for Mapping the Underworld. Geoinformatics, Gaizhuong, China (in prep)

19. Noy, N.F., Musen, M.A., "The PROMPT suite: interactive tools for ontology merging and mapping." International Journal of Human-Computer Studies **59,** 983-1024 (2003)

20. Do, H.H., Rahm, E.: COMA—a system for flexible combination of match algorithms. 28th International Conference on Very Large Data Bases (VLDB) (2002) 610-621

21. Beck, A.R., Fu, G., Cohn, A.G., Bennett, B., Stell, J.G.: Knowledge and Data Integration for Utility Assets: Progress from the MTU and VISTA Projects. UKWIR (in press)

22. Parker, J.: Minimising Street Works Disruption: Buried Asset Data Collection and Exchange Field Trials. UKWIR (2006)

23. Beck, A.R., Fu, G., Cohn, A.G., Bennett, B., Stell, J.G., "Some industry considerations for UK utility data integration." Computers, Environment and Urban System (in press)

24. Belsis, P., Gritzalis, S., Malatras, A., Skourlas, C., Chalaris, I.: Enhancing Knowledge Management Through the Use of GIS and Multimedia. In: Karagiannis, D., Reimer, U. (eds.): Practical Aspects of Knowledge Management. Springer, Berlin (2004) 319-329