MIXHS

Semi Automated Transformation to OWL Formatted Files as an Approach to Data Integration

A Feasibility Study Using Environmental, Disease Register and Primary Care Clinical Data

S. F. Liang[1]; A. Taweel[2]; S. Miles[2]; Y. Kovalchuk[1]; A. Spiridou[1]; B. Barratt[3]; U. Hoang[4]; S. Crichton[4]; B. C. Delaney[1]; C. Wolfe[4]

[1]NIHR Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London, London, UK;
[2]Department of Informatics, King's College London, London, UK;
[3]Environmental Research Group, MRC-PHE Centre for Environment and Health, King's College London, London, UK;
[4]South London Stroke Register, Division of Health and Social Care Research, King's College London, London, UK

Keywords
Informatics, knowledge, semantics, data linkage, OWL ontology

Summary

Introduction: This article is part of the Focus Theme of *Methods of Information in Medicine* on "Managing Interoperability and Complexity in Health Systems".

Background: Data heterogeneity is one of the critical problems in analysing, reusing, sharing or linking datasets. Metadata, whilst adding semantic description to data, adds an additional layer of complexity in the heterogeneity of metadata descriptors themselves. This can be managed by using a predefined model to extract the metadata, but this can reduce the richness of the data extracted.

Objectives: to link the South London Stroke Register (SLSR), the London Air Pollution toolkit (LAP) and the Clinical Practice Research Datalink (CPRD) while transforming data into the Web Ontology Language (OWL) format.

Methods: We used a four-step transformation approach to prepare meta-descriptions, convert data, generate and update meta-classes and generate OWL files. We validated the correctness of the transformed OWL files by issuing queries and assessing results against the original source data.

Results: We have transformed SLSR LAP and CPRD into OWL format. The linked SLSR and CPRD OWL file contains 3644 male and 3551 female patients. The linked SLSR and LAP OWL file shows that there are 17 out of 35 outward postcode areas, where no overlapping data can support further analysis between SLSR and LAP.

Conclusions: Our approach generated a resultant set of transformed OWL formatted files, which are in a query-able format to run individual queries, or can be easily converted into other more suitable formats for further analysis, and the transformation was faithful with no loss or anomalies. Our results have shown that the proposed method provides a promising general approach to address data heterogeneity.

Correspondence to:
Shao Fen Liang
7th Floor, Capital House
42 Weston Street
London, SE1 3QD
United Kingdom
E-mail: Fennie.Liang@kcl.ac.uk

Methods Inf Med 2015; 54: 32 – 40
http://dx.doi.org/10.3414/ME13-02-0029
received: June 21, 2013
accepted: April 23, 2014
Epub ahead of print: June 6, 2014

1. Introduction

In order to deliver a high quality service to patients, integrating data from different healthcare providers to provide a coherent view of an individual's care, and to facilitate healthcare research and quality monitoring has become essential. Healthcare data are complex. And one of the main issues in data integration is that of heterogeneity, physical (data stored in isolated institutions), structural (different source providers have different data models and formats), and semantic (use of different terminologies) [1, 2]. Several approaches have been proposed to overcome the problem of data heterogeneity such as ontology-based approaches [3] and middleware [4].

Ontology-based approaches rely on a foundational ontology, combined with a set of modularised ontologies that model entities and relationships within the data for a specific purpose. Extracted information is transformed into a common format and loaded into targeted databases such as the ontologies created by Tao et al. for extracting heterogeneous data sources to help clinical researchers answering time-related clinical-important questions from EHR [5]. These approaches utilise the modularised ontologies. However, as these are purpose-built, they cannot be generalised to those data outside the pre-built models. Also, this "extract-transform-load" process often discards "unfit" data or degrades the informa-

tion value present in the original data source.

Alternatively, middleware approaches are used for mediating message exchange. These approaches map data into a standard coding system to manage semantic interoperability. Barbarito et al. [6] successfully produced a platform for regional data integration, but re-mapping from their middleware to data sources and training users in managing the platform were cost-intensive and time-consuming tasks.

The Web Ontology Language[a] (OWL) is one of the ontology languages designed for developing and authoring ontologies. It is used for processing information by machine instead of presenting information to humans. OWL is designed to support the Semantic Web activities, which can be used to define classes and properties of those classes, to define individuals and assert properties about them and to reason about these classes and individuals within a domain of interest.

Bouamrane et al. [7] have demonstrated that semantic web technologies can enable sophisticated clinical decision support functionalities. Transforming data into OWL ontologies allowed for explicit expression of all the information implicit in the data and enabled reasoning on the transformed ontologies. Rector et al. [8] also defined a Code Biding Interface in OWL language that successfully mapped HL7 messages to SNOMED CT codes to support healthcare research. We adopted this approach of semantic web technologies for data integration. We hypothesised that transforming data into OWL format without pre-built models would allow different data sources to be linked together in a generic manner and to communicate with existing ontologies without a significant engineering overhead.

We demonstrate the function of this approach by linking air pollution, stroke and primary care clinical data within a geographical area covering south London. Urban air pollution is an environmental health risk, which varies spatially and temporally according to the proximity and strength of pollutant emissions, such as vehicles or industry, and weather conditions. Methods for assessing this spatial-temporal variation within a geographical area are entirely independent from healthcare data, but in order to robustly assess the impact of this health risk on a population, the two datasets must be linked. For example, Atkinson et al. [9] found associations between daily variation in air pollution levels in London and short-term increases in the number of people visiting accident and emergency departments with respiratory complaints. Specific to stroke data, a recent study by Hansell et al. [10] found that areas in west London experiencing high levels of aircraft noise were associated with increased risks of stroke, coronary heart disease, and cardiovascular disease.

2. Objectives

Our primary aim was to link stroke, air pollution and primary care clinical data together to establish an analysis database that would facilitate on going research into linkages between air pollution and stroke over time, across geographical areas and between an expanded range of air pollutants. Therefore, three datasets were chosen for our work: the South London Stroke Register (SLSR), the London Air Pollution toolkit (LAP) and the Medicines and Healthcare Regulatory Authority Clinical Practice Research Datalink (CPRD).

There is inherent heterogeneity between these three datasets. The SLSR and CPRD data are stored in flat files and contain patient-based information, whilst the LAP toolkit is stored in an SQL database and is a geographically-based dataset with very different semantics. Dealing with data heterogeneity from these three datasets is one of the principal challenges of this work.

From the informatics perspective, our objectives were: to avoid the use of predefined models that limit the usage of data from original sources while transforming data into the OWL format; and to maintain data integrity and semantics.

3. The Data Sources

The SLSR [11] is a population-based register, which has recorded data from patients of all age groups for a defined area of South London since 1995 with incident or recurrent strokes. Data are collected at the time of admission, again at three months and one year after initial admission, then annually thereafter. The register monitors stroke incidence rates, recurrence rates, prevalence of stroke-related risk factors, initial impairments, presence of stroke phenotypes and outcomes after stroke, including use of healthcare resources and mortality [12]. This information is collected from each patient. The data are recorded by fieldworkers using a questionnaire and transcribed using SLSR specific codes into values within variable fields. The names of many fields hold semantic information. The field values include a range of numerical values, dates, and encoded strings such as post (ZIP) codes. There are 2195 patients and 76 fields in the SLSR data. ▶Table 1 shows part of the example patients and fields in the SLSR data.

The LAP data contain a single pollutant concentration numerical value for each postcode in London for each year between 2003 and 2010. There are 116,648 records and 17 fields in the LAP data as the example shows in ▶Table 2. Pollution concentrations are calculated using a detailed atmospheric emissions model validated against measured concentrations [13]. This dataset has twelve pollutants, and each assigned a numerical identifier. A metadata table contains information on the model version used to create the pollutant concentrations and the model run date, with one record for each pollutant identifier per model run. This means that concentrations can be subsequently linked back to the underlying model version used. A second metadata table provides a textual description of each pollutant linked to the pollutant identifier.

The clinical dataset from CPRD [14] contains primary care electronic health records, including patients' pseudonomised identification details, practice staff details, type of consultation entered by GP, patients' medical history events, patients' referral and immunisation details from

[a] http://www.w3.org/2004/OWL/

Table 1 Part of the SLSR data example with postcode obfuscated

id	subtype	postcode	sex	employ	migraine	ctime	glas_cs	pri_sbp	pri_dbp	oral6m	bmi_wgt	bmi_hgt	drinker
2249	3	SW9 XXX	1	2	1	930	15	130	80	1	77	1.75	2
2250	2	SE11 XXX	2	5	1	400	15	150	80	3	56	1.57	1
2256	1	SE11 XXX	2	5	1	900	6	158	78	1		1.52	2
2257	1	SW9 XXX	1	1	1	445	15	150	110	1	79	1.77	2
2261	4	SW9 XXX	2	5	1	930	15	120	80	1		1.55	2
2263	4	SE11 XXX	1	5	1	1300	12	118	56	3			2
2267	2	SE1 XXX	2	5	1	900	11	140	85	1		1.67	3
2268	2	SW4 XXX	2	5	1	9999	7	148	86	1			3
2269	2	SE16 XXX	2	5	2	9999	15	156	89	1			1
2270	1	SE16 XXX	1	5	1	2230	12	146	84	1	80		2

Table 2 Part of the LAP data example with postcode obfuscated

Id	RunYear	PostCode	NO	NO2	NOX	O3	OX	PM10	PM10Exhaust-Contribution
2004SE57BX	2004	SE5 XXX	42.23	48.92	91.15	30.48	79.4	26.56	1.48
2005SE41YA	2005	SE4 XXX	30.66	41.16	71.81	35.37	76.53	24.9	0.87
2005SE135BN	2005	SE13 XXX	27.58	39.82	67.39	36.49	76.31	24.9	0.85
2005SE137QU	2005	SE13 XXX	29.7	41.64	71.34	35.5	77.14	25.5	0.98
2006SE279QY	2006	SE27 XXX	27.36	38.97	66.34	38.3	77.28	24.84	0.83
2007SE154AP	2007	SE15 XXX	33.61	44.1	77.7	35.75	79.85	24.76	0.88
2009SE62EG	2009	SE6 XXX	35.95	46.91	82.86	37.56	84.47	23.83	0.95
2009SE166SB	2009	SE16 XXX	31.97	42.4	74.37	37.79	80.19	22.73	0.73
2010SE14XF	2010	SE1 XXX	32.78	43.98	76.76	38.19	82.17	23.78	0.83
2003SE114BE	2003	SE11 XXX	40.93	46.69	87.61	30.3	76.99	30.34	1.51
2003SE85BZ	2003	SE8 XXX	49.31	48.71	98.02	30.27	78.98	31.82	2.04

electronic health records of 8% of the UK population. The structure of the data is complex in that it comprises more than ten tables based on a sample of the clinical CPRD data. Read Clinical Terms Version 2 [15] are used here. The relationships among these tables can be one-to-one, one-to-many or many-to-many. For example, both Patient and Staff tables are in a one-to-one relationship. The other tables can be either one-to-many or many-to-many relations. For example, there may have one patient with many immunisation records, or a single patient has been recorded many times for performing various clinical tasks with different staff. There are 5000 patients in the exemplar CPRD data, where the Patient table contains 21 fields. ▶Table 3 shows part of the example of the Patient table.

Although there is no common field among the three datasets, SLSR shares demographic information with the CPRD Patient and Staff tables, and LAP shares postcode with SLSR. As the three datasets are different in nature, they provide an example of data heterogeneity, and therefore provide a good demonstration dataset for our ontological approach.

4. Methods

We propose a four-step transformation approach for data integration. The first step is to prepare metadata descriptions that describe each field and value of the source data; the second step is to convert semantic meanings back into the values recorded in the data according to the corresponding meta-description; the third step is to gather all the meta-data and to transform them into meta-classes in OWL format. The transformed meta-classes are stored and updated in a meta-class pool; and the last step is to transform source data into OWL files with connections from the meta-classes.

4.1 Step 1: Preparing Semantic Description

This step relies on two inputs: the source dataset and its meta-descriptions. The

Table 3 Part of the Patient table example of the CPRD data

patid	vmid	gender	yob	mob	marital	famnum	chsreg	chsdate	prescr	capsup	ses
471001	9370	2	181	0	6	3871	2		0	4	0
5293001	22496	1	185	0	1	11232	2		0	4	0
5520001	1207	1	187	0	6	538	2		0	4	0
6277001	9947	1	151	0	6	309	2		0	4	0
9469001	15396	2	151	0	4	7683	2		0	4	0
11864001	13633	1	189	0	1	3716	1	14/06/1991	0	4	0
13826001	0	1	182	0	0	11579	2		0	0	0
16256001	0	1	171	0	2	13287	2		0	0	0
19719001	0	2	170	0	2	14989	2		0	0	0
25729001	0	1	206	12	0	15626	1	05/01/2007	0	0	0

meta-descriptions can be either a collection of files or a single file depending on users' convenience. In our case the meta-description of SLSR and LAP are stored as a single file while the meta-descriptions of CPRD are kept as a collection of many small text files in a directory. At this stage of our work, we have restricted the meta-description files to follow a pre-defined format to facilitate the conversion.

These meta-descriptions are essential to add meaning to the structural aspect of the source data. For example, a field titled "eth6cat" can only be understood by the data provider, therefore, a description file is essential to explain that the "eth6cat" means ethnicity, and the value recorded as "1" represents "white", "2" represents "black" and so on. Without the data description, it will be difficult for users to understand the data.

4.2 Step 2: Replacing Codes with their Semantic Descriptions

The conversion process in this step has two advantages; one is to check for any missing meta-descriptions; the other is to enable the subsequent collection of meta-data without the confusion of the different coding systems used by different data providers. For example, one provider may use "F" for "female" and "M" for "male" while another provider uses "1" for "female" and "2" for "male". However, if the semantically meaningful strings, "female" and "male", have been recorded in the data instead of using codes then they will be carried over without any change.

An example of part of the SLSR data before and after the conversion is shown in ▶Table 4 and ▶Table 5. The conversion involved not only the value of the data but also the heading of each column. For instance, "subtype" in ▶Table 4 has been converted into "ocsp (Oxford Community Stroke Classification) classification" in ▶Table 5. This is an important step necessary to model an ontology. Since an ontology contains different entities such as "Classes", "Individuals", "Object Properties" and "Data Properties", we need a mechanism to automatically assign each data value to its correct ontological entity.

This conversion distinguishes between text and numerical values from the original data, so that we can identify Object and Data Properties for our transformation from the data at the next two steps.

4.3 Step 3: Gathering Meta-classes

Since modelling ontologies depends on users' purposes, it is difficult to generalise. However, there are still some common principles for building ontologies that can be applied; an ontology Class should represent more than one Individual; an Object Property maps a relationship between two Individuals; and a Data Property accommodates one specific data value of an Individual.

For our purpose, we transformed each column into a Class, where each unique value contained in this column is the Individual of this Class. For example, if the column "ethnicity" contains values "black" and "white", we then transform "ethnicity" to a Class, and "black" and "white" to Individuals of the "ethnicity" Class.

Table 4 SLSR data before the conversion

id	subtype	sex	employ
2249	3	1	2
2250	2	2	5
2256	1	2	5
2257	1	1	1
2261	4	2	5
2263	4	1	5
2267	2	2	5
2268	2	2	5
2269	2	2	5
2270	1	1	5

These choices depend on the purpose of use and granularity required of the data.

The transformation process employed OWL API[b], and was coded in the Java environment. We used a meaningful naming approach for each generated meta-class. For example, instead of naming a class as "eth6cat" from the original data, the converted data enables the Class to be named as "Ethnicity". This naming approach allows each meta-class to self describe to users.

[b] http://protege.stanford.edu/plugins/owl/api/

unique identifier	ocsp classification	sex	employment status prior to stroke
2249	poci	male	part time
2250	paci	female	retired
2256	taci	female	retired
2257	taci	male	full time
2261	laci	female	retired
2263	laci	male	retired
2267	paci	female	retired
2268	paci	female	retired
2269	paci	female	retired
2270	taci	male	retired

Table 5
SLSR data after the conversion

In this step, if a column contains text values then it is assigned an Object Property while being transformed into a meta-class. So this step will generate a set of Classes, some Individuals of the Classes and some Object Properties. The quantity of resulting meta-classes varies according to how many columns contain text values. Each meta-class, its related Object Property and Individuals are stored in an OWL format, so they can be browsed by any ontology browser [16–19]. Since Protégé is a ubiquitous tool for editing, browsing, reasoning and querying ontology, we used it in this paper for illustrating our work. ▶Figure 1 shows the "Ethnicity" Class example in the Protégé frame.

The axiom for representing the Individual – black african – in OWL format will look like:

```
<owl:NamedIndividual rdf:about="file: Ethnicity.owl#BlackAfrican">
    <rdf:type rdf:resource="file: Ethnicity.owl#Ethnicity"/>
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">black african</rdfs:label>
</owl:NamedIndividual>
```

The generated Object Properties are the key mechanism for maintaining consistent semantic relationships across different ontologies to enable interoperability and linkage between ontologies. For example, both "Staff" and "Patient" ontologies can contain the "ethnicity" Class. Using the same Object Property – "hasEthnicity" – to connect Individuals from "Staff" and "Patient" will enable a reasoner to return a joint count from both "Staff" and "Patient". For example, if the count for the Individual "black african" in "Staff" is 8 and in "Patient" is 12 then we can run a reasoner and issue a Description Logic (DL) [20] query – "hasEthnicity value black african" – to get a returned count as 20.

This step also contains a complex process to update meta-classes. If new terms are detected while transforming newly converted data into our system, these new terms are added as new Individuals to the specified meta-classes. For example, if the gender Class containing two Individuals: "female" and "male" already exist in the meta-class pool from one dataset transformation, we will update it to have a new "Individual "Indeterminate" if this is a new" term to be found from another dataset.

This process only allows new Individuals to be added, but not modified or deleted to affect existing relations. The mechanism for finding which class should be updated is determined by a highest statistic score.

4.4 Step 4: Transforming Source Data into OWL Formatted Files

Our last step is to transform whole data into OWL format. In this transformation, each record containing text values will have its corresponding meta-classes and relative Object Properties applied. Numerical values in the same record are transformed into Data Property values.

For example, if a patient number 2280 has weight 94.5 kg, has height 180 cm, is in white ethnicity and is a male, the transformation result will contain two Data Properties: hasWeightKg and hasHeightCm, two Object Properties: hasEthnicity and hasGender. The relations among those Individuals are:

- Patient number: 2280
- hasWeightKg: 94.5
- hasHeightCm: 180.0
- hasEthnicity: white
- hasGender: male

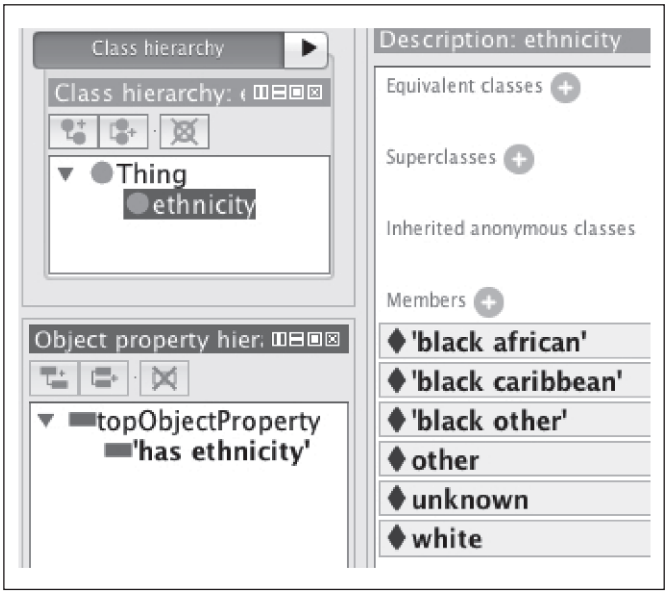

Figure 1
The ethnicity class

Figure 2 Two hundred and thirty-eight Individuals from the SLSR OWL file have been classified into the SNOMED CT SCT_230715005 Class

The purpose of distinguishing Data and Object Properties is that only the numeric values in Data Properties can enable mathematic calculation. For example with an active reasoner if we issue a DL query as "hasWeightKg some double [> 90.0]", we can expect to get all the patient' weight over 90 kg. So the Data Properties are useful for grouping numeric data for a statistical analysis.

The naming approach at this step is different to the meta-class generation. We use a combination of the file name plus the first column heading from the converted data as the name of the OWL file. For example, the resulting "ReferralPatientId.owl" file is generated from a file named "Referral" with "patient id" as the first column heading in the Referral table. This naming approach avoids duplication among meta-classes.

4.5 Link to Existing Ontology

Since ontologies enable reasoning, Classes and Individuals can be implied to be present based on existing relationships. Our generated OWL files provide great flexibility to linking with existing ontologies, such as SNOMED CT [c] and classifications such as ICD10 [d], via the use of common clinical concepts.

We explored connecting our generated SLSR OWL file with SNOMED CT by loading SLSR and SNOMED CT into the same Protégé frame. We found the SNOMED CT term "posterior circulation stroke of uncertain pathology (230715005)[e]" could be used to represent stroke patients whose OCSP classifications were POCI (Posterior Circulation Infarcts). Therefore we added a definition into this Class in Protégé as:

"posterior circulation stroke of uncertain pathology" EquivalentTo "Unique Identifier" and "has ocsp classification" value "poci"

So this Class infers that the "posterior circulation stroke of uncertain pathology" Class in SNOMED CT is equivalent to the "UniqueIdentifier" Class in SLSR and its ocsp classification is "poci", with an active reasoner we got 238 patients from SLSR to be classified into the SNOMED CT (230715005) Class as shown in ▶ Figure 2.

In addition, because the "posterior circulation stroke of uncertain pathology" Class is a sub class of "Stroke of uncertain pathology" we also know that the 238 patients are to be qualified into the "Stroke of uncertain pathology" Class. This kind of connection can be widely applied to a variety of ontologies by our system.

In order to validate our work, we measured the correctness of the transformed OWL files by issuing queries to both individual and combined OWL files and assessing the query results against the original source data.

5. Results

We have built a proof-of-concept system to implement our approach. The transformation process requires users' to select either a directory or a single file of the meta-description and the source data that is to be transformed. We tested the ability of our approach to transform several data sources into OWL files. All of the Object

[c] http://www.ihtsdo.org/snomed-ct/
[d] http://www.who.int/classifications/icd/en/
[e] SNOMED CT code: SCT_230715005

Query with	Query for	Count
SLSR	Male patient	1143
SLSR	Male patient with white ethnicity	716
SLSR	White male patient and had stroke in 2005	94
LAP	NO value bigger than 30 but no more than 40 in 2010	5257
LAP	NO value bigger than 40 but no more than 50 in 2010	893
LAP	NO value bigger than 50 in 2010	360
CPRD	Female patient	2499
CPRD	Female and Single patient	191
CPRD	Female patient and dead before 2010	183

Table 6 Examples of query results from the transformed OWL files

Query with	Query for	Separated Count	Total Count
SLSR	Male patient	2501	3644
CPRD		1143	
SLSR	Female patient	1052	3551
CPRD		2499	
SLSR	Postcode at SE1	354	1752
LAP		1398	
SLSR	Postcode at SE11	172	560
LAP		388	
SLSR	Postcode at SE15	4	1036
LAP		1032	
SLSR	Postcode at SW8	165	663
LAP		498	

Table 7 Query results from two loaded OWL files

Properties and their related Classes of each of the transformed OWL files were imported from the meta-class pool.

The SLSR OWL file contains 41 Classes, 40 Object Properties, 65 Data Properties and 3731 Individuals; the LAP OWL file contains 4 Classes, 3 Object Properties, 14 Data Properties and 119853 Individuals; the CPRD data has been transformed into several OWL files from our collection. For example, the sample patient OWL file contains 10 Classes, 9 Object Properties, 22 Data Properties and 5063 Individuals.

For validating our transformed ontologies, we loaded each of the abovementioned OWL file into Protégé and activated HermiT [21] reasoner. HermiT is one of the efficient reasoners for OWL ontologies. It can determine whether or not the given OWL ontology is consistent; it can classify individuals into classes and identify subsumption relationships between classes, and much more. We then issued three queries to the loaded file. The results have been compared with the original data source to make sure our transformation is not "lossy" and faithful to the original datasets, and the results are shown in ▶Table 6.

In order to confirm that each of our populated ontologies can be integrated with one and another, we loaded two of them into a same Protégé frame then ran a reasoner for issuing DL queries. We first loaded SLSR and CPRD OWL files into a Protégé frame to find out how many male and female patients were present. We then loaded SLSR and LAP into another Protégé frame to obtain patients and air pollution counts in 2010 and classified by outward postcode areas. ▶Table 7 shows some postcodes with the separated and combined counts from SLSR and LAP.

Reasoning over the postcodes also gave us information on five postcodes containing only SLSR but not LAP data, and vice versa for the other twelve postcodes.

▶Table 8 shows zero count from either dataset of the 17 outward postcode areas, where no further data linkage between SLSR and LAP can be done via these postcodes.

6. Discussion

Repurposing data from different heterogeneous sources for use and linkage for other healthcare or clinical research usages is often a very complex and costly process. The novelty of the proposed approach is in its generic method for semantic analysis of data sources and their reusability and linkage across different organisations and with existing third party knowledge bases.

The approach semi-automates the process of data understanding and conversion in a coherent reusable semantic framework.

In our experiment meta-classes containing demographic information such as gender and ethnicity are the most reusable classes, and therefore, they can be the most required elements for linking to other data. For example, both our SLSR and CPRD do not use HL7 standard codes for recording gender. Our Step 2 enables the code "1" and "2" to be converted into "Female" and "Male" where these two terms are coded as "F" and "M" in the HL7 v3 standard. Our conversion step allows the converted semantic meaningful terms to be mapped into HL7 Concept Name, where the data can be linked to other source if HL7 is required. This can be the same approach to apply to some of the ethnicity codes used in HL7 as shown in ▶Figure 3.

Some part of information from the source data may require pre-processing into normalised data such as temporal data in the format as dd-mm-yy, dd/mm/yy, dd-mm-yyyy, or postcode with or with-

out a space between inward and outward areas.

Although few meta-classes can be shared between SLSR, LAP and CPRD, our approach has enabled a further linkage when these data sources are expanded to have more overlapping information. In CPRD data, patient's demographic and staff roles are the most reusable meta-classes. If in the future, for example, as the SLSR gathers care provider information on stroke patients, the current CPRD practice data can be expand to have stroke specific information that would enable the linkage among CPRD, SLSR and LAP to provide advanced information to support further clinical research. In addition, the detailed phenotyping of patients allows us to understand risk of stroke and survival better. So sociodemographic factors may well relate to what is known already about exposure but the details on aetiology of stroke and case severity of stroke is key to identify how air quality affects risk and outcome in different sociodemographic groups, types of stroke mechanism, etc.

Scalability can be a potential problem. Our transformation has increased 14 times the size from a plain text to an OWL file, for example the text file of SLSR data is 668KB but is 9.7MB for the SLSR OWL file. We need at least an 8GB virtual memory machine to run a reasoner to get a quick result. Protégé can take some time to run a reasoner and return query results, and therefore, a better performing user interface will be required for wider use of the method.

Also, because our system is not yet fully automatic, it sometimes requires human interaction to complete a task. The human interaction can become a burden of our users, and it can also lead to errors, therefore, improving automation and reducing human interaction will be our next step to improve the quality of the system.

7. Conclusion

This paper has presented an approach to transform heterogeneous datasets into a

Table 8 Postcodes are not shared by both data

Number	Outward Post-code	LAP	SLSR
1	BR1	252	0
2	BR3	4	0
3	CR4	2	0
4	SE10	23	0
5	SE12	371	0
6	SE13	573	0
7	SE14	339	0
8	SE21	275	0
9	SE22	463	0
10	SE3	138	0
11	SE6	737	0
12	SE7	0	2
13	SE9	35	0
14	SG11	0	1
15	SG16	0	1
16	SW1	0	1
17	SW6	0	1

Figure 3 Ethnicity codes used in HL7 v3 in the Race & Ethnicity group

generic computational format, using semantic web technologies. Our approach enables data to be transformed into a reusable format, while maintaining data integrity and consistency. The generated format captures both the structural semantics and the raw data, which enables querying the data without extensive prior knowledge of its original structure or semantics. This is important, because it enables users to run further analysis on the data without the need to transcribe them and save their time spent on understanding the data.

Our approach generated a resultant set of transformed OWL files, which are in a query-able format to run individual queries, and can be linked with existing ontologies. This approach has been tested on SLSR, LAP and CPRD three different datasets to prove its feasibility and generalisability. Our results have shown that the proposed method provides a promising general approach to address data heterogeneity. Our approach also provides the fundamental step to enable data sources to be linked and queried in various ways to answer more challenging research questions. In the future, we will try to minimise users' need to intervene in the transformation process to improve the quality of data integration.

Acknowledgments

This work is supported by the National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London.
We thank the MHRA CPRD for providing a demonstration copy of the CPRD data and Sean Beevers from King's College London for generating the LAP data.

References

1. Rinner C, Janzek-Hawlat S, Sibinovic S, Duftschmid G. Semantic Validation of Standard-based Electronic Health Record Documents with W3C XML Schema. Methods Inf Med 2010; 49 (3): 271–280.
2. Sachdeva S, Bhalla S. Semantic interoperability in standardized electronic health record databases. Journal of Data and Information Quality 2012; 3 (1): 1–37.
3. Taweel A, Speedie S, Tyson G, Tawil ARH, Peterson K, Delaney BC, editors. Service and Model-driven Dynamic Integration of Health Data. The first international workshop on Managing interoperability and complexity in health systems. Glasgow; 2011.
4. Budgen D, Rigby M, Brereton P, Turner M. A data Integration Broker for healthcare ststems. IEEE Computer 2007; 40940: 34 – 41.
5. Tao C, Pathak J, Welch SR, Bouamrane M-M, Huff SM, Chute CG, editors. Toward Semantic Web based Knowledge Representation and Extraction from Electronic Health Records. Managing Interoperability and Complexity in Health Systems (MIXHS'11). Glasgow, Scotland, UK: October 28, 2011.
6. Barbarito F, Pinciroli F, Mason J, Marceglia S, Mazzola L, Bonacina S. Implementing standards for the interoperability among healthcare providers in the public regionalized Healthcare Information System of the Lombardy Region. Journal of biomedical informatics 2012; 45 (4): 736–745. PubMed PMID: 22285983.
7. Bouamrane M-M, Rector A, Hurrell M. Semi-automatic Generation of a Patient Preoperative Knowledge-Base from a Legacy Clinical Database. OnTheMove (OTM): Berlin Heidelberg: Springer-Verlag. LNCS 5871; 2009. pp 1224 –1237.
8. Rector A, Qamar R, Marley T. Binding Ontologies & Coding systems to Electronic Health Records and Messages. Journal of Applied Ontology 2009; 1: 51–69.
9. Atkinson RW, Anderson HR, Strachan DP, Bland JM, Bremner SA, Ponce de Leon A. Short-term associations between outdoor air pollution and visits to accident and emergency departments in London for respiratory complaints. Eur Respir J 1999; 13 (2): 257–265.
10. Hansell AL, Blangiardo M, Fortunato L, Floud S, Hoogh Kd, Fecht D, et al. Aircraft noise and cardiovascular disease near Heathrow airport in London: small area study. BMJ 2013; 347: f5432: 1–10.
11. Stewart JA, Dundas R, Howard RS, Rudd AG, Wolfe CDA. Ethnic differences in incidence of stroke: prospective study with stroke register. BMJ 1999; 318 (7189): 967–971.
12. Addo J, Bhalla A, Crichton S, Rudd AG, McKevitt C, Wolfe CDA. Provision of acute stroke care and associated factors in a multiethnic population: prospective study with the South London Stroke Register. BMJ 2011; 342: d744.
13. Kelly FJ, Anderson HR, Armstrong B, Atkinson R, Barratt B, Beevers S, et al. The Impact of the Congestion Charging Scheme on Air Quality in London. Part 1. Emissions modelling and analysis of air pollution measurements. Res Rep Health Eff Inst 2011; 155: 5–71.
14. Alexandropoulou K, Vlymen Jv, Reid F, Poullis A, Kang J. Temporal trends of Barrett's oesophagus and gastro-oesophageal reflux and related oesophageal cancer over a 10-year period in England and Wales and associated proton pump inhibitor and H2RA prescriptions: a GPRD study. Eur J Gastroenterol Hepatol 2013; 25 (1): 15 –21.
15. Read JD, Benson TJR. Comprehensive coding. Br J Healthcare Computing 1986; 3: 622– 625.
16. Allemang D, Polikoff I, editors. TopBraid, a multi-user environment for distributed authoring of ontologies. 3rd International Semantic Web Conference (ISWC 2004); Hiroshima, Japan, 2004. Springer Verlag.
17. Kalyanpur A, Parsia B, Sirin E, Grau BC, Hendler J. Swoop: a web ontology editing browser. Journal of Web Semantics 2006; 2 (4): 144 –153.
18. Erdman M, editor Ontology engineering and plug-in development with the NeOn Toolkit. 5th Annual European Semantic Web Conference (ESWC 2008); 2008.
19. Noy NF, Sintek M, Decker S, Crubézy M, Fergerson RW, Musen MA. Creating Semantic Web Contents with Protégé-2000. IEEE INTELLIGENT SYSTEMS: The Semantic Web 2001. pp 60 –71.
20. Baader F, Horrocks I, Sattler U. Description logics as ontology languages for the semantic web. Lecture Notes in Artificial Intelligence 2005; 2605: 228 –248.
21. Motik B, Shearer R, Horrocks I. Hypertableau Reasoning for Description Logics. Journal of Artificial Intelligence Research 2009; 36: 165 –228.

