The Schema Last Approach to Data Fusion

Neil Brittliff and Dharmendra Sharma

University of Canberra

Faculty of Education, Science, Technology and Mathematics

Dharmendra.Sharma@canberra.edu.au

Abstract

Big Data presents new challenges that require new and novel approaches in order to resolve the problems associated with the variability and variety of data obtained from multiple sources. This paper focuses on how to manage variety and the eclectic nature of big data using a technique known as 'Schema Last'. The 'Schema Last' approach is a frame work which defers the application of a descriptive model until it is required. This paper also provides a formal definition of the 'Schema Last' methodology and demonstrates the effectiveness over the more traditional Extract-Transform-Load methodologies employed in many organizations. The 'Schema Last' approach can be used as input to Map Reduction, Index creation and various data mining techniques. Ultimately, the Schema Last approach provides the frame-work to 'fuse' semi-structured data into a single coherent view.

1 Introduction

The Australian Crime Commission (ACC) established the Fusion Data Section in 2010. The section was tasked to collect and analyze data. Legislative powers were granted to the ACC that enables data to be obtained under their coercive powers. These powers do not extend to the format or structure of the data.

The approach taken by the Australian Crime Commission was to model a variety of data sets received via the data collection processes. The ACC could then utilize the collected data for the following purposes:

- assess the behavior of known entities.
- determine any new leads based on the behavior of known entities.
- develop models that may indicate criminal activity within the Australian community.

The *schema last approach* (SLA) is a technique to model data contained within a Big Data store. This approach allows data models to be specified on an on-demand basis and as new knowledge became available, this can be used to respecify the schema definitions. This approach only works if the schema's specification is independent to the data storage. The SLA addresses the issues raised by [Klaus-Dieter Schewe, 2013] in 2012 concerning data quality, interpretation and modification of data as a result of the *cleansing* process.

The size of data sets received by the Australian Crime Commission can range from small to extremely large. Initially, *data cleansing* or the transformation process as part of Extract-Transform-Load (ETL) was the approach taken to convert the raw data into a sanitized form that was loaded into relational tables.

The value proposition for each data source can be quite different. For example, data sets may be classified as low or high signal data sources. Low signal data sources in themselves do not provide any useful indicators but could be used to confirm an entity's address, data of birth and property ownership. An entity membership of low signal data source in itself is not a form of intelligence. High signal data sources can be further analyzed and an entity's membership may indicate a potential threat or unlawful activity that would require further analysis. An example of low and high signal data for attendance at a fairground would be residents of town compared to the entry list of partons.

As the number of data sources continued to increase the ACC came to the realization that the existing ETL process was taking too long and therefore delaying the time taken to analyze the data. Therefore, a new approach had to be found to reduce the demands placed upon the Fusion Data Centre.

1.1 Motivation

The initial implementation utilized a normalized relational table structure. This structure was specifically designed to capture the following *entity* information that included:

- person: name, date-of-birth, and address details.
- organization: organization name, address, Australian business number (ABN), and Australian Company Number (ACN).
- information pertaining to the relationship between two or more entities.

Any data item that did not fall within the structure was discarded and therefore the model could only capture simple identity information. In addition, transaction, time series data or highly linked data could not be captured within this structure even though most relational database products allow for the modification of existing table structures.

There were instances where data definitions within the schema did not adequately describe the source data. In this case a person within the Fusion Data

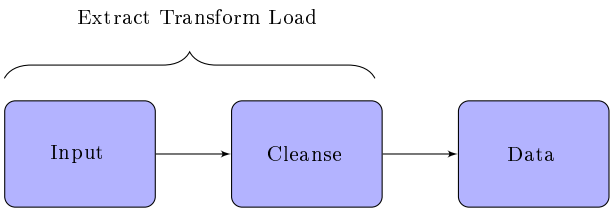

Figure 1: Data Cleansing

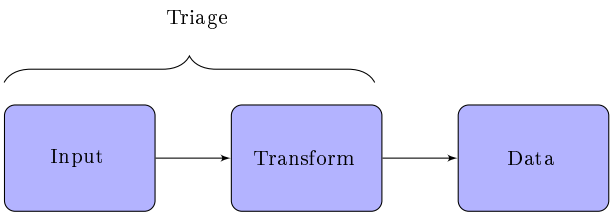

Figure 2: Data Triage

Centre would determine the most appropriate field within the relational structure and perhaps alter the raw data item to comply to the field's schema definition.

This whole entire data ingestion process was seen by the management of the fusion data centre as cumbersome and error prone. Therefore an alternate solution had to found to replace the existing ingestion process.

1.2 Data Triage and ETL

Data Cleansing is the transformation of data from a non-canonical to a canonical state. ETL involves the transformation of data into a state suitable for ingestion into a database. This usually requires the standardization of data fields from the original source to a new target format.

For example dates may be transformed from **mm/dd/yyyy** to **dd/mm/yyyy** or addresses may be required to have the postcode field removed. Extraneous attributes within a field or the entire field are lost due to the cleansing process. For example a person's salutation (MR, MRS, DR, *et cetera*) if contained within a field may be required to be removed to comply to the name field specification. Often the ETL process can lose data or perhaps pervert the original data in some way and may in turn reduce overall quality of the data [Rajaraman, 2014]. As argued by N. Brierley, T. Tippetts and P. Cawley:

> "Formal data cleansing can easily overwhelm any human or perhaps the computing capacity of an organization." [N. Brierley and Cawley, 2014.]

This problem was also identified by Vincent Burner in 2007:

> "that the data volume may overwhelm the Extract Transform Load process and that *data cleansing* may introduce unintentional errors." McBurney [2007]

Data Triage is a different approach to ETL in that the raw value of the data is always maintained throughout the transformation process. The data is loaded into the data stores **verbatim** unless there are structural transformation issues with the original data source.

To summarize the differences between the Data Triage and ETL:

- Data Triage does not alter the original data value or format whilst ETL may alter a field's content.
- Data Triage will not eliminate any data field contained within the original data source whilst the result of an ETL load process will eliminate fields that do not comply to a fixed schema.

The purpose to cleanse or not cleanse can best be expressed utilizing the **Beliefs**, **Desires** and **Intentions** [Bratman, 1999] methodology based upon Michael Bratman's theory of human practical reasoning [Castanedo, 2011]. Michael Bratman's theory can be applied to data analysis where the data modeler takes the following into consideration:

Beliefs Beliefs provide the *inference rules* to process and manipulate the incoming data. These rules can be captured and reused for new data sets or reapplied to existing data sets.

Desires Desires represent how the data can best be used or processed. Desires represent the motivation behind the data and are generally expressed as an accomplishment. For example, this data can be used to determine if this person is what is commonly expressed as 'on the move' or 'up to no good'.

Intentions Intentions represent the deliberate use of the data and how the data is to be used. For example, the data may obtained so that it correctly enhances the intelligence surrounding a particular criminal organization.

However, there is a cost to data cleansing:

- Inconsistent processing where cleansing involves human assessment.
- The elimination or removal of unwanted tokens within a field can lead to the introduction of errors and inconsistencies.
- The introduction of human judgment which in turn may lead to errant assumptions that result in decisions based on false or misleading data interpretation.

Jimmy Lin and Dmitiry Ryaboy state:

> "That a major problem for the data scientist is to *flatten the bumps* as a result of the heterogeneity of data." [Lin and Ryaboy, 2013]

1.3 Data Provenance

Data provenance is the ability to retain the lineage between the original and processed data. This is an important attribute of data quality as Paulo Pinheiro da Silva states:

> "without acknowledging data provenance the quality of data will decline." [Paulo Pinheiro da Silva, 2007]

Paul Groth, co-chair W3C Provenance Working Group said:

> "The provenance of information is crucial in deciding whether information is to be trusted, how it should be integrated with other diverse information sources, and how to give credit to its originators when reusing it,.." [Oracle - Release, 2013]

2 Data Integration and Matching

Many challenges face all Big Data implementations with the ability to combine eclectic data sources into a single view being the most significant. Over time, as new knowledge about the data source becomes apparent, this in turn may lead to a greater understanding of the data structure. This is where SLA allows one to adapt the model to conform to a new understanding of the data. To facilitate the integration process a universal schema can be applied to all data sources which in turn can be used to provide a *common* view to the data sources. It is important characteristic of SLA that it does not change any individual data values.

3 Data Fusion and the Schema

Data fusion is the process where data sources are combined into a single view. This view can be used as input to various analytical and data-mining techniques and still utilize the SLA schema from the ingestion process.

The schema may change over time and this will not affect the raw data the schema represents. A SLA schema can be used for the following purposes:

- Provide a consistant view of the data source contained within the *big data* repository.
- Assist in the creation indexes and formulate index strategies.
- Identify portions of data source suitable for extraction and further analysis.
- An input **map** to a *Hadoop map reduction* task.

3.1 Schema First Frameworks

The history of database implementations including the *codaysl* network model, hierarchical database systems such as Information Management System (IMS) by IBM and more recent relational database implementation require a schema definition to be established before any data can be stored. Changes can be made to the schema afterward but will ultimately alter the data represented by the schema. For example, an addition of a column to a schema within a relational table will result in a *null* value for every row contained within that table. If a column is removed from the schema then all the associated data for the deleted column is lost. The issue was identified by IBM in 2011:

> "The schema is not meta-data but *combines* the data with a structural representation. Therefore, changes to the schema results in changes to the data even if it is only the addition of a *null* field value to each record." [IBM, 2011].

Schema protocols like Thrift, Avro, CORBA, DCOM and XML all have a schema definition language and fall into the category of schema first whereby the schema is fixed and difficult to change after inception.

Figure 3: Schema Last Models

3.2 The Schema Last Model

The model is a logical group of fields. A model may hold fields contained within another model. Models that do not hold fields from other models are said to be **distinct** (figure 4). Models containing all the fields held within another model are said to be encapsulated (figure 6). There is no restriction on the number of encapsulated or distinct models contained within a SLA schema. Models may share fields within the same schema definition (figure 5).

The same model definition may reappear in another data source schema. It is also possible that two or more data sources have identical SLA schema definitions.

3.3 The Field Label

Each field within the SLA schema definition must have an associated label. The default value for the label should be the name given to the field within the original data source. If there is no name within the original data source then a descriptive name is assigned to the label.

3.4 The Field Domain

Each field may have an associated domain where the domain determines the field's potential range of values. A typical domain would be **name** where the **name** may contain either a person or organization name. If the field were only to contain only a person name then **person-name** would be the field's domain. Other domains may be: phone-number, complete-address, contact-address. It is important to restrict or manage the number of domains and remove any ambiguity amongst domain specifications. A domain should not be confused with a primitive data type (string, integer, float, *et cetera*) and a primitive data type is not descriptive enough to be used as a domain.

3.5 Ontological Support

Not many data modeling products have Ontology support [Maxim, 2013], however W3C has specified Ontology Web Language (OWL) which defines a comprehensive ontological language [W3C, 2012]. OWL allows the definition of hierarchical ontological structures based on the RDF specification. SLA can also assist in the ontological description of the data in regards to domain classification as shown in figure 8

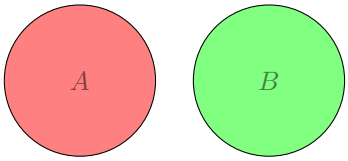

Figure 4: Distinct Models

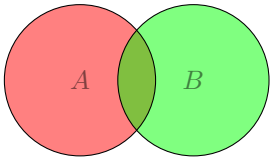

Figure 5: Overlapping Models

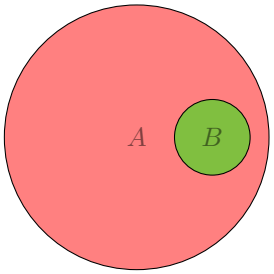

Figure 6: Encapsulated Model

and OWL could be used to describe the ontological structure of the domain.

4 Formal Description

SLA provides a formal *framework* to describe the process and the language necessary to describe a data source. The process provides an inherent feedback loop and is an iterative process. It is always possible to 'start over again' if the SLA schema definition is proved to be incorrect.

The SLA process consists of the following phases:

1. **Data source triage**: The coercion of data into a form where it can be uploaded into the data repository. No data is lost or changed during the **triage** process.
2. **Schema specification**:
   (a) The labeling of all fields contained within a data source.
   (b) The association of each field with a specific domain.
   (c) The creation of models and associated field memberships.
   (d) Schema storage and version management.
3. **Suitability**: Determine the schema is an accurate representation of the data source.
4. **Application**: Create indexes based on the schema defined in step 2.
5. **Verification**: Ensure that by application of the Schema that erroneous indexes are not created and that the models defined within the Schema correspond with the data contained within the data source.
6. **Fuse**: Identify entities which are common within the data source.
7. **Resolve**: The construction of a *'single source of truth'* to represent an entity (generally a person or organization) within the entire data repository.

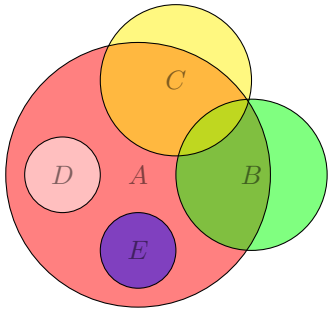

Figure 7: Multiple Encapsulated Models

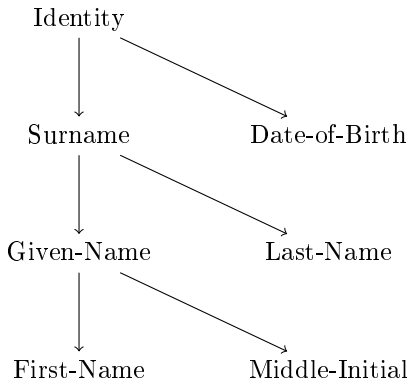

Figure 8: Ontological Support

The Schema definition is formally represented in RDF N3 form[W3C Group, 2014]. RDF blank nodes are used to group the fields that are contained within a model. Figure 11 is a simple schema definition containing three fields and two models.

4.1 Schema Storage Considerations

There is no prescriptive persistent storage technology and there is an advantage to store the Schema with the raw data.The resultant removal of the raw data will lead to the destruction of the data's associated Schema. The Minerva reference implementation stores the data as an RDF list structure. The Schema is then stored within the same RDF graph alongside the RDF list.

4.2 Map Reduction

Map Reduction is a strategy that has become popular to process large data-sets. To summarize, the *map* phase maps the input data into a known structure and the reduction phase summarizes the data. Map Reduction has become a tool of choice amongst data mining practitioners [Rajaraman, 2014]. The Map Reduction approach can take full advantage of the SLA in that the SLA schema can be used as the input map.

5 Data Fusion

Data Fusion is the process of integration of multiple data data sources into a single view. Federico Castanedo defines data fusion as:

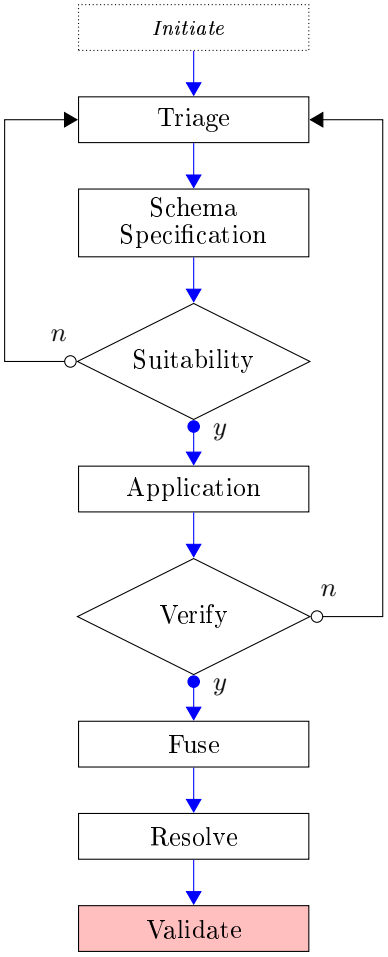

Figure 9: Schema Last Process

> "The integration of data and knowledge from several sources is known as data fusion." [Castanedo, 2013]

The SLA provides the framework to achieve this, in that a consistent modeling technique is applied to all the data sources contained within the repository. The *fused* data can then be modeled again and the result in itself can become another data source. The Data Fusion consolidation provides a unique perspective to the Big Data repository. It is essential to refer back to the original raw data source in order that the correct data provenance is maintained.

5.1 Indexing Strategies

Indexes are a crucial part of any Big Data implementation and it is important that the indexes are independent to the stored data. The SLA schema defines the models and structure of the indexes that will be used to query the data contained within the repository. The indexes can be tailored to a specific purpose for example: an index may be a person's name and allow for typographical errors (see figure 12). [Rajaraman, 2014, Christen, 2012].

The SLA Schema can be used as the basis to generate any number of index strategies. For example, SOLR which is a document index system based on Lucene is a high performance text based indexing engine. SOLR indexing definitions can utilize the SLA schema that contains sufficient meta-data to define SOLR filters and tokenizers.

```
@prefix rdf: <http://www.w3.org/...#>
@prefix rdfs: <http://www.w3.org/...#>
@prefix schema: <http://www.sla.org.au/...#>

<file:/schema.ttl> schema#:schema
 [
 <schema#://model>
   [
    rdf:#_1 _:b1 ;
    rdf:#_2 _:b2 ;
    rdf:_#3 _:b0
  ];
 <schema#://model>
  [ rdf#:_0
   _:b0 , _:b2 , _:b1 ;
    rdf#:_1
   _:b0 , _:b2
  ]
 ] .
_:b0 rdfs#:domain "document-id" ;
   rdfs#:label "document-name" .
_:b1 rdfs#:domain "first-name" ;
 rdfs#:label "given-name" .
_:b2 rdfs#:domain "last-name" ;
 rdfs#:label "surname" .
```

Figure 10: Schema Last Definition in N3 format

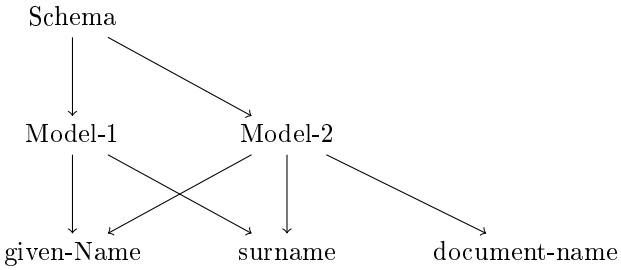

Figure 11: Schema Definition - Visual Representation of figure 10

5.2 Classifiers and Stochastic Attributes

A row may contain one or more models and a model can represent an entity that may be an individual or company [Rajaraman, 2014]. Each data source could be assigned additional attributes or classifiers in the form of *meta-data*. Classifiers are used to describe the nature of the data source and how the data source was originally obtained. Additional attributes would also add value to the data source; for example: Geo-spatial coordinates, time of ingestion and an intelligence ratings.

5.3 Meta-data

Additional meta-data in the form of name/value pairs or tags are used to annotate the data source or models contained within the data source (see figure 13). This includes:

- the security classification (Secret, Highly Protected, Unclassified),
- the name of the organization that supplied the data
- any special data handling requirements.

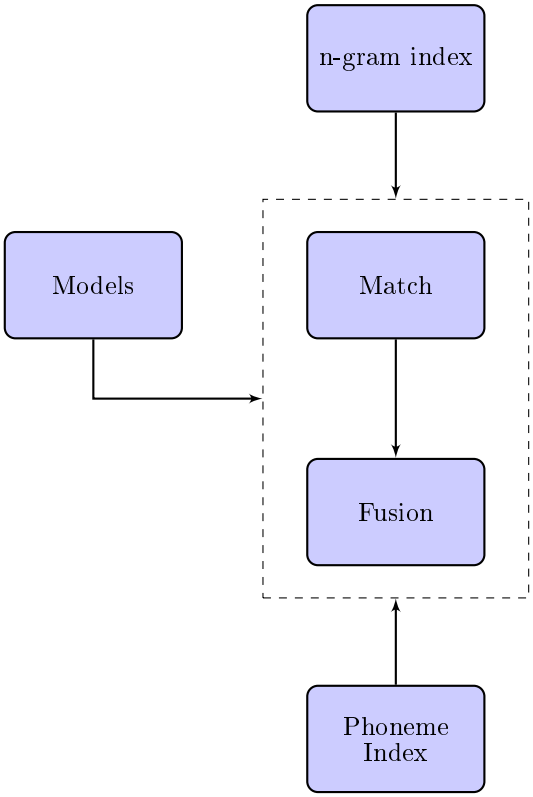

Figure 12: Index Strategies and Schema Last

Meta-data were also used to describe any associations amongst data sources within the data repository. For example, if two data sources are related in some way then this relationship could be captured within descriptive meta-data tags.

It is essential to conform to, or at the very least map to, known standards for meta-data and that the meta-data is consitent across all data sources, rather than using proprietary or homegrown schemes. In addition the meta-data scheme(s) are defined within a thesaurus that has been developed for the provision of a common vocabulary across the data sources.

> "Good meta-data conforms to community standards in a way that is appropriate to the materials in the collection, users of the collection, and current and potential future uses of the collection." [NISO, 2007]

5.4 The Single Source of Truth

Part of the Identity Resolution process is to formulate the *golden record* which is a consolidated view of an individual or organization. The consolidated view in turn can contain a *match* score between the two matched entities. It is important that the new knowledge that pertains to each record can be reapplied in the construction of the consolidated view. The consolidated record can be queried by the user and this will contain all the current information relevant to the entity.

6 Initial Findings

The 'Schema Last' approach to data fusion has been in operation within the ACC since 2012 and during that time there has been over a thousand data sources processed that utilized this technique. Initial results have shown a dramatic reduction in the time taken to

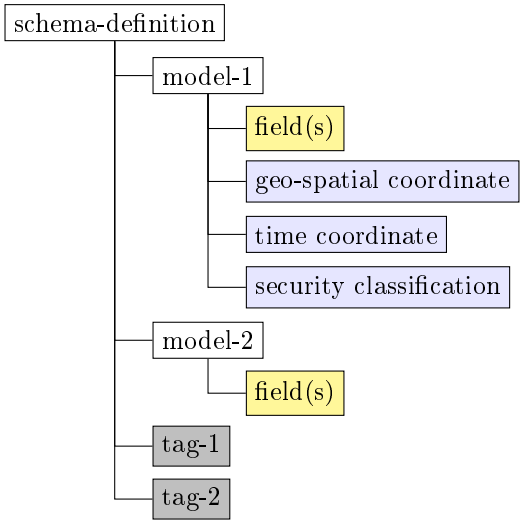

Figure 13: Model Description including Attributes and Classifiers

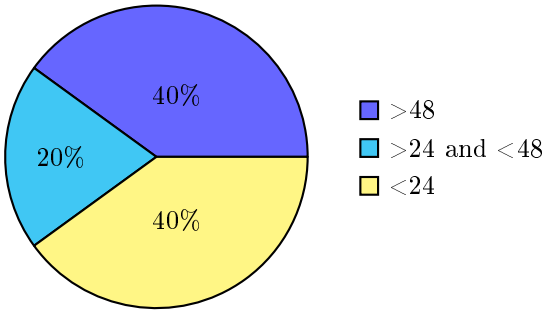

Figure 14: Time taken in hours to cleanse 50 Data Sources

collate data sets received by the ACC. One particular data set took over three months to cleanse; whilst utilizing SLA this particular data set only took minutes to upload and model. The reaction by the analysts within the ACC has been positive and they have commented on how vital information was lost due to the previous ETL process. SLA has allowed the data source modeling process to be delegated to unskilled staff which has allowed the skilled data analysts to focus on model development and target detection. The improvements are shown in both figure 14 and figure 15.

In addition, it took only **thirty** domains to describe every field contained within the data sources.

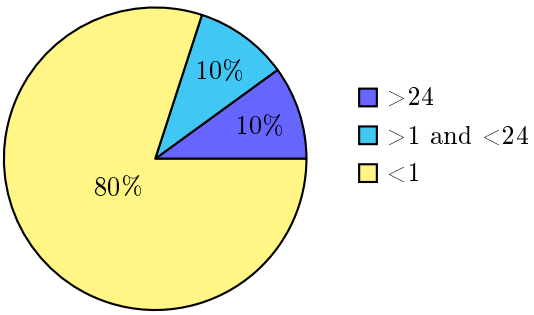

Figure 15: Time taken in hours to triage 1010 Data Sources

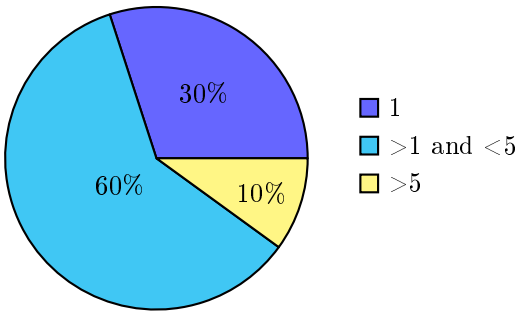

Figure 16: The number models required to describe a data source

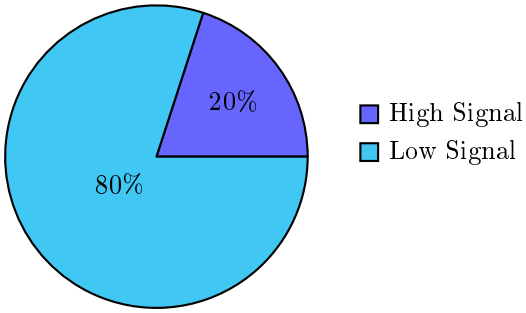

Figure 17: The Ratio of High to Low Data Sources in terms of number

The domains types were deliberately restricted. For example, the domain *document identifier* was used to describe any unique identifier within a data source. There were no detectable overlaps as a result of this generalized classification of domains.

The Minerva Schema Designer 19 was developed to allow users to create and modify SLA Schema Definitions. The schema models could be saved along side the data or separately as a file. The interface was deliberately kept simple to minimize user training requirements and reduce modeling errors.

The reaction from users of the application was positive. Other Australian government departments have shown interest with the *schema last approach* and have expressed a desire to run a proof-of-concept within their respective organizations.

6.1 Future Work

The resultant 'fused' data has raised a number of concerns among the data analysts, their greatest concern is the ever increasing number of results returned from

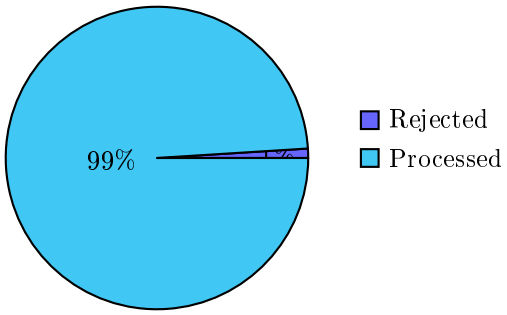

Figure 18: The ratio of processed to rejected data sources

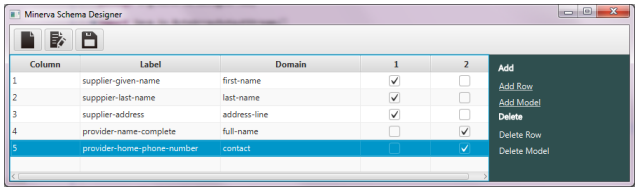

Figure 19: Minerva Schema Designer

search queries. To determine a *single source of truth* for an entity has proved to be a challenge especially when there is no unique identifier to link one entity across multiple data sources. The 'schema last approach' is first step towards this goal and work can be done in this area to improve the accuracy in establishing links between *like* entities within different data sources.

7 Conclusion

The ACC now has a tool to address the eclectic nature of the data sent to them. SLA can be used to define index strategies, provide the *map* in Map Reduction and the form foundation for data mining. The formal definition of the schema syntax allows for the **interchange** of models and the **sharing** of meta-data among organizations and institutions.

SLA has provided the platform to *fuse* data into a consolidated view and to resolve issues associated with variability and variety of data when obtained from multiple sources.

References

Michael E. Bratman. ***Beliefs, Desires and Intentions.*** **CSLI Publications, 1999.**

Federico Castanedo. A multi-agent architecture based on the bdi model for data fusion in visual sensor networks. ***Journal of Intelligent & Robotic Systems*****, 62(3-4):299–328, 2011. ISSN 0921-0296. doi: 10.1007/s10846-010-9448-1. URL** `http://dx.doi.org/10.1007/s10846-010-9448-1`**.**

Federico Castanedo. A review of data fusion techniques, 2013. URL `http://dx.doi.org/10.1155/2013/704504`**.**

Peter Christen. ***Data Matching.*** **Springer, 2012.**

Paulo Pinheiro da Silva. Propagation and provenance of probabilistic and interval uncertainty in cyberinfrastructure-related data processing and data fusion. ***DEPARTMENTAL TECHNICAL REPORTS (CS)*****, (UTEP-CS-07-56), 11 2007. URL** `http://digitalcommons.utep.edu/cgi/viewcontent.cgi?article=1224&context=cs_techrep`**.**

W3C Working Group. Resource description framework (rdf), 2 2014. URL `http://www.w3.org/RDF/`**.**

IBM. Terminology: Dynamic- vs. fixed-schema databases, July 2011. URL `http://www.dbms2.com/2011/07/31/dynamic-fixed-schema-databases/`**.**

Qing Wang Klaus-Dieter Schewe. Knowledge-aware identity services. *Knowledge and information systems*, **36:335–357, 2013.**

Jimmy Lin and Dmitriy Ryaboy. Scaling big data mining infrastructure: The twitter experience. *SIGKDD Explor. Newsl.*, **14(2): 6–19, April 2013. ISSN 1931-0145. doi: 10.1145/2481244.2481247. URL** http://doi.acm.org/10.1145/2481244.2481247.

Bakaev Maxim. Ontology to support web design activities in e-commerce software development processmore. **2013. URL** http://www.academia.edu/929051/Ontology_to_Support_Web_Design_Activities_in_E-Commerce_Software_Development_Process.

Vincent McBurney. 17 mistakes that etl designers make with very large data, **2007. URL** http://it.toolbox.com/blogs/infosphere. **http://it.toolbox.com/blogs/infosphere/17-mistakes-that-etl-designers-make-with-very-large-data-19264.**

T. Tippetts N. Brierley and P. Cawley. Data fusion for automated non-destructive inspection. *Proceedings of the RSPA*, **2014. URL** http://rspa.royalsocietypublishing.org/content/470/2167/20140167.abstract.

NISO. A framework of guidance for building good digital collections, **2007. URL** http://www.niso.org/publications/rp/framework3.pdf.

Anand Rajaraman. *Mining of Massive Datasets.* **2014. URL** http://infolab.stanford.edu/~ullman/mmds/book.pdf.

Oracle Press Release. Oracle implements w3cs standard for data provenance. **2013. URL** http://www.oracle.com/us/corporate/press/2028860.

W3C. Owl 2 web ontology language structural specification and functional-style syntax, **December 2012. URL** http://www.w3.org/TR/owl2-syntax/.