# An Ontology-Based Data Integration Framework for Construction Information Management

Akinyemi, Abiodun Gideon; Sun, Ming; Gray, Alasdair J. G.

*Published in:*
Proceedings of the ICE - Management, Procurement and Law

*DOI:*
10.1680/jmapl.17.00052

*Publication date:*
2018

*Document Version*
Peer reviewed version

*Citation for published version (APA):*
Akinyemi, A. G., Sun, M., & Gray, A. J. G. (2018). An Ontology-Based Data Integration Framework for Construction Information Management. Proceedings of the ICE - Management, Procurement and Law, 171(3), 111-125. DOI: 10.1680/jmapl.17.00052

**Submitted:** 09 October 2017

**Published online in 'accepted manuscript' format:** 14 March 2018

**Manuscript title:** An Ontology-Based Data Integration Framework for Construction Information Management

**Authors:** Abiodun Akinyemi[1], Ming Sun[1], Alasdair J. G. Gray[2]

**Affiliations:** [1]School of Energy, Geoscience, Infrastructure and Society, Heriot-Watt University, Edinburgh, UK. [2]School of Mathematical & Computer Sciences, Heriot-Watt University, Edinburgh, UK.

**Corresponding author:** Abiodun Akinyemi, Heriot-Watt University, School of the Built Environment, Edinburgh, EH14 4AS, United Kingdom. Tel.: +447518548754

**E-mail:** aga1@hw.ac.uk

## Abstract

Information management during the construction phase of a built asset involves multiple stakeholders using multiple software applications to generate and store data. This is problematic as data comes in different forms and is labour intensive to piece together. Existing solutions to this problem are predominantly in proprietary applications, which are sometimes cost prohibitive for small engineering firms; or conceptual studies with use cases that cannot be easily adapted. In view of these limitations, this research presents an ontology-based data integration framework that makes use of open source tools that support Semantic Web technologies. The proposed framework enables: rapid answering of queries over construction data integrated from heterogeneous sources; data quality checks; and reuse of project software resources. The attributes and functionalities of the proposed solution align with the requirements common to small firms with limited information technology skill and budget. Consequently, this solution can be of great benefit for their data projects.

## 1. Introduction

A study by McKinsey Global Institute revealed that productivity in the global infrastructure sector has remained static at 1% for the past 20 years (Barbosa et al., 2017). The lack of progress is, in part, due to the fragmented nature of the construction industry (Farmer, 2016). Fragmentation in the construction process impacts decision making because project stakeholders usually: (i) have limited interactions (Mohd Nawi, Baluch and Bahauddin, 2014); and (ii) generate construction information based on individual work requirements (Fiatech, 2011). This problem calls for data integration, a service that combines disparate data and outputs coherent information for decision support (Doan et al., 2012). An analysis of data from industrial construction projects by Zhai et al. (2009) showed that construction labour productivity improves with the use of automation and integration technologies. For large construction organisations with ample resources, this challenge is addressed with investment in tools that use big data technologies like Hadoop, HBase, Neo4J etc (Bilal, Oyedele, Akinade, et al., 2016). However, for small firms, such capital expenditure is not usually possible. Instead, they often seek ways to avoid incurring costs in handling data existing in heterogeneous formats (Walsh, 2016).

Given the benefits of digital technologies to industrial work processes in general (Cordes and Stacey, 2017), several studies in recent times (Bilal, Oyedele, Qadir, et al., 2016), have focused on exploring how technology can improve construction work processes. One of the technologies that have found useful applications in the construction industry is the Semantic Web (Abanda, Tah and Keivani, 2013). However, very few studies exist on demonstrating use cases and implementation plans that can be easily adapted (Pauwels et al., 2017). This study fills this knowledge gap by demonstrating the value of a proposed technology framework implemented with open source software applications that support Semantic Web technologies. This will bring the benefits of Semantic Web technologies closer and cheaply to non-programmers working on complex data tasks during construction. Table 1 provides definition of some of the key terms used in this paper.

## 2. Theoretical foundation

This study uses the Technology-Organisation-Environment (TOE) framework (DePietro, Wiarda and Fleischer, 1990) as the theoretical foundation for investigating appropriate data integration method for construction information management. According to DePietro, Wiarda and Fleischer (1990), the TOE framework is an Information Systems theory that focuses on the process by which a firm adopts and implements technological innovations. The framework is established on three tenets - technology, organisation and the environment. The technological context is focused on available technologies and their relevant characteristics to a firm; organisational context concerns internal factors within an organisation (e.g. budget, work process,  structure); and environmental context covers factors outside the organisation (e.g. stakeholders, regulations, market structure) (Baker, 2012). Technological context in this study is about the availability of data integration software applications that can deliver the desired construction data quality; the organizational context is about the limited resources of small businesses supporting construction activities; and the environmental context refers to the fragmented nature of the construction industry. These three elements are essential in the way small firms needing to integrate construction data search for and adopt appropriate technologies. It is within this scope that this study proposes a technology framework for data integration during construction.

## 3. Research methodology

Figure 1 shows the research design for this study. The first step was to define the research problem and identify a theoretical framework for the study as described above. Following this, a construction stakeholder, Ariosh (www.ariosh.com), was consulted to identify a data integration scenario for which the organisation has no tool and is unwilling to assign a budget for a commercial tool (Step 2). By interviewing the engineers involved with the data integration task in the organisation, the appropriate requirements were elicited and used to define the acceptance criteria for proposed solution (Step 3). A systematic literature review (Okoli and Schabram, 2010) on (i) construction data qualities; (ii) common data integration techniques in construction; (iii) ontology application in construction and (iv) adoption of open source software by organisations was thereafter carried out (Step 4). The requirements elicited were mapped to appropriate technologies identified from literature review considering the constraint of limited budget (Step 5). As a result, appropriate open source software applications that can be easily accessed by the targeted group were researched and used for design implementation. Using Rapid Application Development (RAD) technique of software development (Despa, 2014), a data integration framework is proposed (Step 6) and demonstrated with a hypothetical example based on the real practice of the consulted construction stakeholder (Steps 7 & 8). According to Despa (2014), RAD Is a minimalist and agile software development method that limits planning and focuses prototyping on using reusable components. The demonstration case is evaluated using tests that indicate the acceptance criteria (Step 9). Also, it is worth noting that the authors' perspective to this research is that of a systems analyst (Misic and Graf, 2004).

## 4. Overview of construction data integration

### *4.1 Data quality dimensions*

Construction is a collaborative activity involving a multi-disciplinary team, including client, architects, engineers, consultants, contractors, etc (Niknam and Karshenas, 2015). Each member of this team is responsible for some aspects of the project and often relies on information produced by others. Fiatech (2011) suggests that where all construction data are supplied in a single technology supplier system, the process of combining data and extracting information is somewhat straightforward. However, because of the multi-stakeholder nature of construction, some flexibility is required with technology use to accommodate the domain requirements of each stakeholder (Omar and Nehdi, 2016). Consequently, relevant data during construction sometimes do not combine as required - because of data quality issues; and often need reprocessing to facilitate integration (Soibelman et al., 2008). To maximise the value from construction data, its qualities in relation to the construction process must be taken into consideration. Figure 2 captures the key construction data quality dimensions as discussed in Westin and Sein (2014), and Bilal, Oyedele, Qadir, et al. (2016). These dimensions are described as follows:

a) *Volume***:** No one set of construction data is likely to be large enough to become unmanageable. However, the total amount of data generated over the construction cycle can be large enough that it can become challenging to sieve out useful information for the operations team at handover (Yang, 2009). Fiatech (2011) describes handover process as arduous and prone to error as data size grows because of the difficulty in aligning and making connections between multiple data sources.

b) *Variety***:** Construction activities generate significant amounts of data in a wide variety of formats (Sun and Aouad, 1999; Mutis and Issa, 2012). Text documents, videos and pictures are plentiful during construction (Soibelman et al., 2008). Also, the various software applications used during construction produce files in a variety of syntaxes e.g. Extensible Markup Language (XML) in Yurchyshyna and Zarli (2009), and comma-separated values (CSV) in Bilal, Oyedele, Akinade, et al. (2016). Variety also relates to the variations that exist when multiple sources represent the same real-world entity differently, regardless of identical export formats.

c) *Veracity***:** This is about information integrity and focuses on the accuracy and currency of data (Kitchin, 2014). It is important that data being used for tasks during construction represents what they are intended to represent. When they do not, errors are introduced into tasks. Simple things like using an old report for a decision or planning based on inadequate drawing detail can have significant risk and cost implications. Veracity also covers managing null values, misleading values, outliers and non-standardised values that may be present in data (Rubin, 2014).

d) *Velocity*: In construction, this is about the need for regular updates to project data - underscored by Omar and Nehdi (2016). Progress reports and schedule data fall in this category. As there is always the need to have the most accurate and up-to-date information in a timely manner, the frequency of updates requires consideration in managing construction data.

The discussed issues are a replica of those associated with 'Big Data' by Kitchin (2014) except for the issue of very large data volumes. Regarding data volume, Westin and Sein (2014) indicated that irrespective of the size of the organisation, investments must be made in information technology resources to ensure efficient handling of very large volumes of data. Usually, small firms act as subcontractors and handle smaller subsets of construction data; so, issues other than data size are pertinent. Consequently, Big Data technologies seem appropriate in addressing the data integration challenges in relation to highlighted construction data quality dimensions. Applying Big Data technology in construction could mean investment in proprietary Big Data tools and skilled personnel, or developing own solutions using open source software. Bilal, Oyedele, Qadir, et al. (2016) argued that the inability to do either of these efficiently and cost effectively has the potential to adversely impact a project. An industry research on the business value of data (Friedman and Smith, 2011) underscores the importance of data quality by stating that 40% of the anticipated value of all business initiatives are not achieved due to the poor quality of data. This is most likely true because poor data quality has efficiency implications, increases risks and compromises decision making (Zhang et al., 2018). In the current construction environment, high-quality data are a requirement for successful data-driven projects (Westin and Sein, 2014). However, considering the low profit margin of construction jobs reported by Bilal, Oyedele, Qadir, et al. (2016), acquisition of costly technological solutions is not a first choice for small businesses (Dainty et al., 2017).

### *4.2 Common data integration techniques*

According to Mutis and Issa (2012), integrating data from multiple sources within a construction project can be labour intensive and often incurs a high cost in terms of expert resources. This is because every software package has a unique way to structures data entities and label their properties – called data model. This data model may not be open, thus making automatic data integration difficult. Although technologies like XML and relational database schemas can handle the syntactic differences between structured heterogeneous sources, these technologies cannot resolve the semantic incompatibilities without a careful mapping effort (Wiesner et al., 2011). The commonly used data integration techniques in the construction industry include:

a) *Document-based information integration*: In this method of data integration, there is no knowledge of the content of the source files other than the metadata attributed to them. The metadata is also the basis for the storage of the sources in a database. Given that most construction data are available in propriety electronic formats, they can be easily managed in this way. However, this integration method is shallow and requires further exploration of the sources after document retrieval. An example is when all the documents related to a project are labelled with a unique code and stored in a database.

This integration method will allow such documents to be retrieved with queries that connect them via the unique code. However, each retrieved document must be read with appropriate software application to find out if it has relevant contents. Electronic Document Management Systems (EDMS) are an example of this technique (Qady and Kandil, 2012).

b) *Data Mashup***:** In this method, the output data is not truly integrated as the source data are only juxtaposed for human interpretation. The function that executes this is coded into a software application with access to the data sources as shown in Figure 3(i). An example is the merging of 3D point cloud from laser scanning with 3D geometric model in the graphical user interface of an application for clash detection or design verification (Son, Bosché and Kim, 2015).

c) *Schema-based integration***:** This type of data integration technique ingests data in its native format and reconciles it to the data model of the host application as explained in Doan et al. (2012). As shown in Figure 3(ii), this process involves mapping data from multiple heterogeneous sources to the schema of a host application. The schema of the receiving application is the mediated schema. An example of this in construction is combining Computer Aided Design (CAD) models from two different software applications by using a third application.

d) *Model based Integration***:** This type of integration as described in Sun and Aouad (1999), is the usual configuration in Enterprise Management Systems (EMS) with a unified repository. The integration process has logical understanding of the transactions between the applications and services within the confederation; and allows full and logical data integration. For example, if there is a rule that application2 in Figure 3(iii) requires input from application1 for a task to proceed, then a missing input from application1 will freeze the process.

### *4.2 Ontology-based data integration*

#### 4.3.1 Ontologies

According to Gruber and Borst (2009),

> "An ontology is a formal, explicit specification of a shared conceptualization. Conceptualization refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine-readable. Shared reflects that notion that an ontology captures consensual knowledge; that is, it is not private of some individual but accepted by a group".

An ontology contains classes, relations, attributes, formal axioms, functions and instances. Classes are used to represent concepts that are either physical or abstract. When a class is derived from another class, it is called a subclass; and the class from which it is derived, called superclass. Relations are used to represent association between the concepts; attributes are used to describe the features of the concepts; formal axioms are used to make assertions about the ontology; functions are used for special cases of relations; and instances represent individuals within the ontology (Bermejo, 2007). Ontology types include (i) top level and (ii) domain specific ontologies. Top level ontologies use generic concepts for their entities (Guarino and

Oberle, 2009) while domain specific ontologies make use of concepts that are common to users of a specific discipline e.g. oil and gas industry in Fiatech (2011). Ontology-based data integration makes use of ontologies in the integration of disparate data. The ontologies, according to Doan et al. (2012), serve as the mediated schema and data sources are described in terms of their entities.

### 4.3.2 Semantic web

The Semantic Web is the manifestation of Tim Berners-Lee's aspiration (Berners-Lee, Hendler and Lassila, 2001) of a web that is capable of automatic processing of data with minimal interaction with people. To achieve this, Doan, Halevy and Ives (2012) explains that semantic markup is being associated with the content of the web to enable easier integration of heterogeneous data, and more accurate search results. To facilitate markup on the web, Resource Description Framework (RDF) (Cyganiak, Wood and Lanthaler, 2014), RDF Schema (RDFS) (Brickley and Guha, 2014) and Web Ontology Language (OWL) (Motik et al., 2012) were developed. These languages according to Doan et al. (2012) are based on the principles of knowledge representation languages which are grounded in description logic, a subset of first-order logic; and are used for defining ontologies.

Cyganiak, Wood and Lanthaler (2014) describes RDF as a directed, labelled graph made up of sets of triples – subject, predicate and object. It is a self-describing data representation language that supports several vocabularies at the same time and has several valid serializations (e.g. RDF/XML, Turtle, N-Triples) (Curé and Blin, 2015). RDF data is queried using SPARQL, a World Wide Web Consortium (W3C) standard (The W3C SPARQL Working

Group, 2013). SPARQL queries RDF data by matching the graph patterns between the queries and the data. RDF triples include International Resource Identifiers (IRIs), which provide a mechanism for referring to resources in a global way, like the Uniform Resource Locators (URLs) described in Fiatech (2011) used on the World Wide Web. This will ensure that the resource is available to anyone wishing to use it. A concern with the use of IRIs stated in Doan et al. (2012), is that they result in lengthy names for resources. Consequently, qualified names, which represent shorter names for the prefixes of resources, are used.

According to Brickley and Guha (2014), RDFS is used for modelling the classes, hierarchies and class membership of individuals in an ontology. Class refers to the nature of things that allows them to be grouped according to some criteria (ISO, 2003). RDFS is also used for specifying the domain and range of relationships in an ontology. All the constructs of the RDFS are contained in OWL. OWL in addition, has several constructs that improve its expressive power, enabling it to model more complex and incomplete domains. Doan et al. (2012) states that OWL was developed to facilitate inferencing over the Semantic Web and Motik et al. (2012) attributes the trade-off that exists between its expressive power and ability to perform inference for its multiple profiles.

The Semantic Web allows rules to be applied to ontologies modelled with RDFS and OWL (Horrocks and Patel-Schneider, 2011). These rules are defined using Rule Interface Format (RIF) (Bruijn and Welty, 2013) and they enable the discovery of extra information in the ontologies by generating new relationships based on existing ones. Concisely, the relevance of Semantic Web in ontology-based data integration includes: its use for developing the ontologies used as mediated schema; its use in describing and linking data from heterogenous sources, making them machine readable; and lastly, its supports for querying and reasoning about the integrated data.

### 4.3.3 Construction applications

Pauwels et al. (2017) observed that the application of the Semantic Web and ontologies for data integration in construction, though not new, is not common. Most of the available examples are conceptual studies and not simple enough for regular engineers to quickly adopt, thus leaving a gap between theory and practice (Zhang et al., 2018). Construction processes that have had ontology-based data integration proposals include planning, monitoring, controlling and compliance checking (Table 2). For construction planning, Zhong et al. (2015) introduced an ontological approach for developing and verifying construction technical plans, and Zhang et al. (2018) presented an ontology approach to support smarter planning decisions on construction cost, environmental impact and safety. Ontology-based methods have also been developed for construction risk planning (Ding et al., 2016) and cost estimation (Abanda et al., 2011). In respect of the monitoring and controlling of construction projects, proposals for code checking have been presented by Yurchyshyna and Zarli (2009) and Zhang and Issa (2011). In addition, proposals on performance checking (Pauwels, Deursen, Verstraeten, et al., 2011) and defect management (Park, Lee and Kwon, 2013) have been demonstrated. Other recommendations have focused on 3D model data integration (Pauwels, Deursen, Roo, et al., 2011) and general construction data integration using ontologies and the Semantic Web (Elghamrawy et al., 2009).

## 5. Adoption of open source software by organisations

Open source software is developed in a public collaborative manner and shared with a license that allows the users to reuse, modify and redistribute them without limitations (Marsan et al., 2012). Open source software is also free and can be easily downloaded from the Internet (Dabbish et al., 2012). As many businesses seek ways to avoid licensing fees of commercial software, open source software seems a viable alternative (Nagy, Yassin and Bhattacherjee, 2010). According to Hauge et al. (2010), adoption of these products by organisations can be either by directly applying them in their operations or by using the products in software development. They identified software development activities to include: making use of open source software development platforms, extending a software application, integrating software applications, adopting open source software development practices, and participating in open source software development communities. The two scenarios that are applicable in this research include using open source software as-is and integrating open source software applications into a system. An important benefit of adopting open source software is that no special skill or experience is needed to install the software or integrate it into specified applications. Challenges that may arise relate to the cost of learning and understanding the software components, and the time it takes to integrate them (Chen et al., 2008). However, these issues are mitigated by adopting open source software with active community support and useful documentation (Sarrab and Rehman, 2014). Overall, Hauge et al. (2010) concluded that the availability of open source software allows businesses to adopt new technologies faster, increase innovation, and improve on productivity.

## 6. Proposed data integration framework for construction information management

### *6.1 Requirements*

Several studies (Wiesner et al. (2011); Mutis and Issa (2012); and Bilal, Oyedele, Qadir, et al. (2016)) show evidence of the need for data integration in construction projects and justify the variety of solutions currently available to the industry. However, small businesses are unlikely to rush to invest in information technology resources for data integration (Horakova et al., 2013). Instead, existing evidence by Walsh (2016) show that these businesses encourage their personnel to explore learning programming languages to develop in-house tools or surf the web for open source tools that can handle their needs. This evidence was validated by interviews with engineers in Ariosh, a medium engineering construction contractor in West Africa. The following requirements were gathered from the elicitation process:

a) *Data migration***:** Effort and time required to extract data from different sources are factors in the choice of tools or technology framework for data integration. A recent survey of data scientists indicated that they spend 60% of their time cleaning and organising data and nearly 60% of them consider this exercise the least enjoyable of their tasks (CrowdFlower, 2016). Consequently, there is a desire for a straightforward process that allows data in the sources to be transformed in a reasonable amount of time.

b) *Data quality checks***:** Being able to check for inconsistencies of the integrated data is very important to the technology users because they do not want to have spent a lot of effort integrating data to then find out in the end that there are errors. As a result, the introduction of quality check measures in the data processing cycle is considered valuable in any tool or technology framework that enables data integration.

c) *Reuse and Extendibility*: Serious care is required for every data integration task to avoid errors as it takes time to complete. Consequently, any solution that allows reuse of data mapping and query files multiple times is very desirable. Also, the possibility of extending the integrated data without errors is considered as valuable.

### *6.2 Conceptual framework*

The proposed data integration framework is based on the consideration that engineers are not computer scientists but capable of understanding and making use of technical knowledge. While engineers may not be able to build comprehensive or perfect ontologies, they can create functional ontologies that will serve their information management purposes. Figure 4 shows the conceptual framework for the ontology-based work process for data integration in construction. The steps in the process are discussed below:

a) *Data Collation***:** Soibelman et al (2008) lists some of the typical data sources in construction. These sources include structured data (e.g. relational databases and spreadsheets); semi-structured documents (e.g. CSV and XML); unstructured text documents (e.g. contracts, specifications, change orders, requests for information, meeting minutes, e-mail messages, webpages); and other unstructured multimedia data (e.g. 2D/3D drawings, pictures, audios, videos). With data extraction techniques (Rusu et al., 2013), these sources can be transformed into preferred structured formats using open source tools (Groves, 2016; Roman et al., 2017). However, the data cleaning effort required depends on the status of the source, which is highly variable. Given the variability of the sources, this research is not focused on the data cleaning process. It is premised on the notion that most of the sources can be exported or organised into a spreadsheet document. Spreadsheets are favoured for this type of data integration task because of their widespread use in engineering (Lee et al., 2016); and multiple open source tools, e.g., MappingMaster (O'Connor et al., 2010) and dot15926 (TechInvestLab.ru, 2013), support mappings between them and ontologies. This step is very important because the spreadsheet templates generated will be repeatedly used for updates on a project. In addition, they can be used for similar projects.

b) *Ontology Development*: Individuals developing ontologies often have different conceptualisations of a given domain, and may all be right, as there is no single right way of modelling a domain. Consequently, it makes sense that for a small data project, an ontology based on the understanding of the data sources by the users is what is developed. Also, it is easier to interpret, accommodate and model what is than trying to fit reality to a standard ontology (Rezgui et al., 2011). To develop required ontology, a list of entities in the sources is first generated. Thereafter, classes are identified and sorted into taxonomy of superclasses and subclasses where applicable. Attributes and relations are also identified. An ontology of the entities in the data sources, as documented in spreadsheets, is then created, and axioms defined. Open source tools like Protégé (Musen, 2015) and Jena Ontology API (Carroll et al., 2004) are recommended for this.

c) *Data Mapping***:** This is carried out to resolve the terminology heterogeneity in the data sources. Once the data sources are organised into spreadsheets and the project ontology developed, mappings are created to migrate the data in the spreadsheets into the ontology. The output data is the integrated project data. Mappings used can be generated using RDF Mapping Language (RML) (Dimou et al., 2014) or any open source library that supports mapping from spreadsheets into OWL e.g. MappingMaster. An important advantage of this step is that it lends itself to self-checks. Any errors in the integrated data can be easily corrected by revising the mapping rules and re-importing the source data into the base ontology.

## 7. Demonstration case study

Pipe spools are common components in large oil and gas construction projects. Activities involved in pipe spool fabrication include cutting, fitting, welding, quality inspection, non-destructive testing, post heat treatment and painting (Standards Norway, 1996). In the fabrication of pipe spools, raw pipes and pipe fittings (e.g. elbows, flanges, reducers) are welded according to a design. To achieve this, piping isometrics (iso) from the design are divided into spools that can be fabricated. These sub-assemblies are tacked, welded and checked for quality of work. Additionally, x-ray tests are carried out to certify the welds, and leak tests completed to check for fluid leakage. This demonstration case is adopted from a pipe spool fabrication project, conducted by the consulted contractor, that was having challenges with progress information.

Information on the progress of work exists in multiple records and is making the determination of the overall completion status of each spool piece inconclusive. Information sources, as shown in Figure 5, include a line list export from a 3D design application (iso number, service, line size); spool status report from the project engineer (iso number, Spool number, completion status); leak test report from a service contractor (Type, spool number, completion status); and x-ray test report from another service contractor (spool number, completion status). The x-ray of welded joints precedes pressure tests in the work process, and reports from the test activities are used to update the project engineer's overall project status.

A problem with the data is that there are gaps in two of the reports – overall and x-ray. The affected reports do not represent the most up-to-date situation at the fabrication yard because some of the spools have had to be retested. Another problem is that stakeholders have represented iso numbers differently e.g. '18"-P-A12-002 001-B' in the pressure test report is represented as '18"-P-A12-002 001B' in the x-ray report and split into '18"-P-A12-002 001' and 'B' in the overall project report. In carrying out a gap analysis of the spools, it is important to use all the reports to resolve the differences, a process that needs to be repeated throughout the duration of the project. On a typical fabrication project, there could be thousands of spools (Soleimanifar, 2016) and it will be very inefficient to manually retrieve the correct status of each spool by looking through each report. As a result, for such an ad hoc data project, this ontology-based data integration work process is appropriate.

*7.1 Application of proposed framework*

To address the above stated problems, the data sources are pre-processed and an ontology of all the available data is created. Thereafter, available data from the sources are mapped to the appropriate entities in the ontology. By doing this, a uniform representation of the data sources is generated, and queries can be used to check for consistency and extract information of interest - including the most up-to-date status of each spool piece. Following the steps of the conceptual framework, the first task is to collate the data. The pre-processed data are as presented in Figure 6 and include annotations identifying the source of each data point recorded under the header 'Issue'. Also, iso numbers in the overall project report are fully completed to ensure each row has a value.

The second step is to create the ontology for the domain. The generated ontology shown in Figure 7 is developed using the earlier mentioned ontology design tool, Protégé. The development process involves inspecting the data sources and identifying the concepts within them. This reveals that multiple stakeholders are involved in the reporting process. Also, concepts in the domain include spool, service and pressure test types. By combining these, it can be deduced that spools in the domain have completion statuses from all the reports; undergo some type of pressure test; and share size and number with documented services. The latter assertion is represented in Figure 7 through a blank node described by Cyganiak, Wood and Lanthaler (2014) as an anonymous resource used for creating links between classes. The air leak and hydro tests are classified as subclasses of pressure test. In OWL, classes are assumed to overlap. Consequently, it is important to make classes sharing a superclass, e.g.

‘HydroTest’ and ‘AirLeakTest’, disjoint from one another. Applying disjoint classes will ensure that an individual cannot be an instance of more than one of the classes. Also required for the data integration process are the assertions about relationships between concepts. For example, ‘Spool is a Thing that has completion progress status for pressure test, x-ray test and the overall project; and the OWL datatype property for Spool is literal string’. These assertions are illustrated in Figure 7.

The third and final step is to map the prepared data in the spreadsheets to the created ontology. This is done using the MappingMaster plugin to Protégé called Cellfie (Kaur and Aggarwal, 2017). An important consideration in the mapping process is ensuring the correct OWL constructs are used in mapping the data. For Cellfie, the transformation rules generated are based on OWL Manchester syntax (Horridge and Patel-Schneider, 2012). Figure 8 shows the transformation rules for the data sources in this case study. The mapping for x-ray report, for example, includes axioms describing elements in column A (Figure 6) as OWL individuals having OWL type ‘Spool’. These individuals have OWL fact ‘hasXrayStatus’, a spool property shown in Figure 7, with values in column B of the x-ray report (Figure 6). The last rule for mapping x-ray report simply adds a comment about the source of entities in the mapping, stating that the latter individuals have a source with value in column C of the x-ray report (Figure 6). This described technique is used to map all the spreadsheet sources. In a situation where there is a relation between entities in different sources (spreadsheets), the appropriate reference notation for the open source tool should be explored. An example in this case study is the mapping rules for pressure test report, shown in Figure 8. The fourth line of the transformation rule describes elements in column B of pressure test report (Figure 6) as OWL individuals that are same with elements in column A of x-ray report (Figure 6). Once the transformation rules for the sources are completed, data from the different spreadsheets can be added to the ontology to generate an integrated data.

### *7.2 Information retrieval*

Having semantically enriched, searchable data can significantly improve the process of information extraction (Shayeganfar et al., 2009). In construction, according to Bilal, Oyedele, Qadir, et al. (2016), data collected are frequently consulted for decision making. Similarly, the RDF data generated from this case study will support decision making. To ask questions of this data, SPARQL queries can be easily parsed via open source graphical user interface (GUI) tools like Twinkle (Almendros-Jiménez, Becerra-Terón and Cuzzocrea, 2017) and YASGUI (Rietveld and Hoekstra, 2014). However, in order to demonstrate the benefit of adding rules to data, the integrated data in this case study was deposited in an open source triple store called Eclipse RDF4J (Knap et al., 2016). Eclipse RDF4J is an open source Java framework for storing, inferencing and querying RDF data. The following query exercises demonstrate information retrieval with and without inferencing.

a) *Query without Inference:* The first two queries in Figure 9, Q1 and Q2, extract distinct instances of spools and the progress statuses recorded for each of them in results R1 and R2. Q1 simply says select distinct spools, their x-ray statuses, their pressure test statuses and their statuses in the overall report where RDF triples about them match RDF triples listed after the 'where' clause. In Q2, inspection of the sources shows that a spool is being represented differently across reports. As a result, ?y and ?spool are used to identify the versions. Q2 simply says select distinct spool (?spool version), their x-ray statuses, their pressure test statuses and their statuses in the overall report where RDF triples about them match RDF triples listed after the 'where' clause. These views reveal the inconsistencies present in the current record and creates opportunity for resolving them.  Q3 counts the number of spools in the project to be 6.

b) *Query with Inference*: An advantage of using the Sematic Web is that the technology has the capability to reason logically about the project data and the rules added to it (Horrocks and Patel-Schneider, 2011). In the latter results, Q1 and Q2 extracted the correct documented statuses of spools in the different reports. However, there is still a need to mentally resolve the logical inconsistencies in the output. For example, according to adopted work process, it is inconsistent for a spool to be undergoing pressure test when it has not gone through x-ray test. To resolve this issue, rules are added to the repository and queries with inference, as shown in Figure 10, are run. SPARQL Inferencing Notation (SPIN) (Knublauch, 2013) is used to represent the rules. SPIN is supported by RDFJ4, so the rules are added to the triple store in addition to the project data. Rule 1 simply asks that every spool with pressure test status as 'Ongoing' should have x-ray status as 'Completed' and overall project status as 'Ongoing'. Pressure test report is used as the benchmark because it is the only report with up-to-date status. Consequently, when it has a status 'ongoing', x-ray status should be 'Completed' and overall project status should duplicate its status. Rule 2 follows a similar construction, and query Q4 returns the true status of every spool in every report.

*7.3 Evaluation*

Acceptance criteria for evaluating the proposed framework are derived from the requirements elicited from the construction stakeholder (see section 6.1). The simple assessment rule is that the proposed solution should give positive indication on each criterion. The following assessments were carried out:

a) *Time:* Figure 11 compares the time expended in using the proposed approach with the time taken to search through the reports manually and determine the true status of a spool. With a good understanding of all the necessary concepts and tools, it takes an hour to generate the case study project data, rules and queries. On the other hand, it takes about five minutes to look through all the reports and determine what the correct statuses of a spool should be. On projection, by the time this manual process is repeated 15 times, the upfront cost of developing the ontological solution would have been fully covered - because it takes about a minute to modify the query for any given spool and get results. Considering there are thousands of spools on large oil and gas fabrication projects according to Soleimanifar (2016), by the time the task on determining the statuses of a spool in the different reports is carried out 200 times (for example), the ontology-based solution would only have taken 259 minutes compared to 1000 minutes in a manual process. This is an outright 70% saving, which will grow further as such task continues.

b) *Data quality checks:* The use of SPIN rules in the proposed solution allows for automatic detection of data inconsistencies - something that is not possible in the manual inspection approach. In Figure 9, even though the data in the sources have been integrated and retrieved correctly, the output data is not logical. For example, it is not possible for spool 18-P-A12-0020001B to have a project status 'Ongoing' when the x-ray and pressure test statuses are 'Not Started'. This is so because adopted work process requires x-ray activities to precede pressure testing, and the pressure test report is used for populating the overall project report. By adding rules, as in the proposed method, inconsistencies are eliminated, and the correct result is outputted as in Figure 10.

c) *Reuse and Extendibility:* Typically, reports are updated periodically on a project. In such a scenario, the mappings, queries and rules will remain the same for the ontology-based framework. The updated data in different sources will only need to be reimported into the initial ontology. This eliminates the need for any reconciliation that may be required in the manual approach. Another advantage of the proposed framework is that people having similar work process will not need to start from scratch. They may only need to adjust an existing template which will further reduce the setup cost. If the same template is used across multiple projects, the saving on the setup cost will further multiply. And when a template is well tested, it will become an organisation process asset over time.

## 8. Conclusion

This study shows that leveraging data integration technologies for construction tasks can be beneficial. It reveals that by applying open source tools based on Semantic Web technologies, it is possible to make significant time savings compared with the traditional manual data management approach. The approach demonstrated in this paper is also low cost, responding to the need of small firms lacking the budget for proprietary technology tools. As demonstrated by the case study example, data in varying forms, originating from multiple sources, can be easily integrated; and the integrity of the output data will remain consistent with the original sources. Also, the mappings, rules, queries and designed ontology can be revised and reused multiple times. While there is an initial setup cost with the proposed approach, demonstrated prototype indicates such cost is quickly recovered within just a few tasks and significant saving is achieved as the number of tasks increases. While the result of the proposed framework indicates potential productivity gains for data projects, its full potential needs to be evaluated by more substantive construction case applications. Consequently, for the next step, this framework will be applied by engineers working on similar tasks in construction and an evaluation on productivity, scalability and cost benefits will be carried out. Future research effort will also seek to further reduce the setup cost of the proposed framework to lower the barrier for its application.

## Acknowledgement

The case study example in this paper is adapted from the fabrication assurance work process of Ariosh, an engineering and construction service company in West Africa. The opinions,

findings and recommendations in this work are those of the authors and do not necessarily reflect the views of the company.

## References

Abanda, F. H., Tah, J. H. M. and Keivani, R. (2013) 'Trends in built environment Semantic Web applications: Where are we today?', Expert Systems with Applications. Pergamon, 40(14), pp. 5563–5577. doi: 10.1016/j.eswa.2013.04.027.

Abanda, F. H., Tah, J. H. M., Pettang, C. and Manjia, M. B. (2011) 'An ontology-driven building construction labour cost estimation in Cameroon', Electronic Journal of Information Technology in Construction, 16, pp. 617–634.

Almendros-Jiménez, J. M., Becerra-Terón, A. and Cuzzocrea, A. (2017) 'Syntactic and semantic validation of SPARQL queries', in SAC '17 Proceedings of the Symposium on Applied Computing. Morocco, pp. 349–352.

Baker, J. (2012) 'The Technology–Organization–Environment Framework', in Dwivedi, Y., Wade, M., and Schneberger, S. (eds) Information Systems Theory. Integrated. New York, NY: Springer.

Barbosa, F., Woetzel, J., Mischke, J., Joao Ribeirinho, M. and Sridhar, M. (2017) Reinventing construction through a productivity revolution. Available at: http://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/reinventing-construction-through-a-productivity-revolution (accessed 06/02/2018).

Bermejo, J. (2007) A simplified guide to create an ontology. Available at: http://tierra.aslab.upm.es/documents/controlled/ASLAB-R-2007-004.pdf (accessed 06/02/2018).

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) 'The Semantic Web', Scientific American, pp. 29–37.

Bilal, M., Oyedele, L. O., Akinade, O. O., Ajayi, S. O., Alaka, H. A., Owolabi, H. A., Qadir, J., Pasha, M. and Bello, S. A. (2016) 'Big data architecture for construction waste analytics (CWA): A conceptual framework', Journal of Building Engineering, 6, pp. 144–156. doi: 10.1016/j.jobe.2016.03.002.

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A. and Pasha, M. (2016) 'Big Data in the construction industry: A review of present status, opportunities, and future trends', Advanced Engineering Informatics. Elsevier Ltd, 30(3), pp. 500–521. doi: 10.1016/j.aei.2016.07.001.

Brickley, D. and Guha, R. V. (2014) RDF Schema 1.1, W3C Recommendation. Available at: https://www.w3.org/TR/rdf-schema/ (accessed 06/02/2018).

Bruijn, J. de and Welty, C. (2013) RIF RDF and OWL Compatibility (Second Edition), W3C Recommendation. Available at: https://www.w3.org/TR/2013/REC-rif-rdf-owl-20130205/ (accessed 06/02/2018).

Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A. and Wilkinson, K. (2004) 'Jena: implementing the semantic web recommendations', in Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters. New York, pp. 74–83. doi: 10.1145/1013367.1013381.

Chen, W., Li, J., Ma, J., Conradi, R., Ji, J. and Liu, C. (2008) 'An empirical study on software development with open source components in the chinese software industry', in Software Process: Improvement and Practice, pp. 89–100. doi: 10.1002/spip.361.

Cordes, F. and Stacey, N. (2017) Is UK Industry ready for the Fourth Industrial Revolution? Available at: https://media-publications.bcg.com/Is-UK-Industry-Ready-for-the-Fourth-Industrial-Revolution.pdf (accessed 06/02/2018).

CrowdFlower (2016) CrowdFlower Data Science Report. Available at: http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf (accessed 06/02/2018).

Curé, O. and Blin, G. (2015) RDF database systems: triples storage and SPARQL query processing. Elsevier.

Cyganiak, R., Wood, D. and Lanthaler, M. (2014) RDF 1.1 Concepts and Abstract Syntax, W3C Recommendation. Available at: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/ (accessed 06/02/2018).

Dabbish, L., Stuart, C., Tsay, J. and Herbsleb, J. (2012) 'Social coding in GitHub: transparency and collaboration in an open software repository', in Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp. 1277–1286. doi: 10.1145/2145204.2145396.

Dainty, A., Leiringer, R., Fernie, S. and Harty, C. (2017) 'BIM and the small construction firm: a critical perspective', Building Research and Information. Taylor & Francis, 45(6), pp. 696–709. doi: 10.1080/09613218.2017.1293940.

DePietro, R., Wiarda, E. and Fleischer, M. (1990) 'The context for change: Organization, technology and environment', in Tornatzky, L. G. and Fleischer, M. (eds) The processes of technological innovation. Lexington, MA.: Lexington Books, pp. 151–175.

Despa, M. L. (2014) 'Comparative Study on Software Development Methodologies', Database Systems Journal, 5(3), pp. 37–56. doi: 10.1109/MAHC.1983.10102.

Dimou, A., Sande, M. Vander, Colpaert, P., Verborgh, R., Mannens, E. and Van De Walle, R. (2014) 'RML: A generic language for integrated RDF mappings of heterogeneous data', in Proc. of the 7th LDOW workshop.

Ding, L. Y., Zhong, B. T., Wu, S. and Luo, H. B. (2016) 'Construction risk knowledge management in BIM using ontology and semantic web technology', Safety Science. Elsevier, 87, pp. 202–213. doi: 10.1016/J.SSCI.2016.04.008.

Doan, A., Halevy, A. and Ives, Z. (2012) Principles of Data Integration. Elsevier.

Elghamrawy, T., Boukamp, F. and Kim, H.-S. (2009) 'Ontology-based, semi-automatic framework for storing and retrieving on-site construction problem information – An RFID-Based Case Study', in Construction Research Congress, pp. 457–466.

Farmer, M. (2016) The Farmer Review of the UK Construction Labour Model: Modernise or Die. Available at: http://www.constructionleadershipcouncil.co.uk/news/farmerreport/ (accessed 06/02/2018).

Fiatech (2011) An Introduction to ISO 15926. Available at: http://www.fiatech.org/images/stories/techprojects/project_deliverables/iso-intro-ver1.pdf (accessed 06/02/2018).

Friedman, T. and Smith, M. (2011) Measuring the Business Value of Data Quality, Gartner. Available at: https://www.data.com/export/sites/data/common/assets/ (accessed 06/02/2018).

Groves, A. (2016) ‘Beyond Excel: how to start cleaning data with OpenRefine’, Multimedia Information and Technology, 42(2), pp. 18–22.

Guarino, N. and Oberle, D. (2009) Handbook on Ontologies. doi: 10.1007/978-3-540-92673-3.

Hauge, Ø., Ayala, C. and Conradi, R. (2010) ‘Adoption of open source software in software-intensive organizations – A systematic literature review’, Information and Software Technology. Elsevier, 52(11), pp. 1133–1154. doi: 10.1016/J.INFSOF.2010.05.008.

Horakova, M., Skalska, H., Olszak, C. M., Ziemba, E., Fitriana, R., Djatna, T., Kursan, I., Mihić, M., Albescu, F., Pugna, I. and Paraschiv, D. (2013) ‘Business Intelligence and Implementation in a Small Enterprise.’, Journal of Systems Integration, 4(2), pp. 50–62. doi: 18042724.

Horridge, M. and Patel-Schneider, P. F. (2012) OWL 2 Web Ontology Language Manchester Syntax, W3C Working Group Note. Available at: https://www.w3.org/TR/owl2-manchester-syntax/ (accessed 06/02/2018).

Horrocks, I. and Patel-Schneider, P. (2011) ‘Knowledge Representation and Reasoning on the Semantic Web: OWL’, in Handbook of Semantic Web Technologies. Springer, pp. 365–398.

ISO (2003) Industrial automation systems and integration — Integration of life-cycle data for process plants including oil and gas production facilities — Part 2: Data model. First edit. Geneva. doi: 10.1109/IEEESTD.2007.4288250.

Kaur, N. and Aggarwal, H. (2017) ‘Evaluation of Information Retrieval Based Ontology

Development Editors for Semantic Web', International Journal of Modern Education and Computer Science, 9(7), pp. 63–73.

Kitchin, R. (2014) 'Big data', in The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences, pp. 67–79.

Knap, T., Hanecák, P., Klímek, J., Mader, C. and Necaský, M. (2016) 'UnifiedViews: An ETL Tool for RDF Data Management', Semantic-Web-Journal.Net, 0(1), pp. 1–15.

Knublauch, H. (2013) SPIN - SPARQL Syntax, W3C Member Submission. Available at: http://spinrdf.org/sp.html (accessed 25/09/2017).

Lee, D.-Y., Chi, H., Wang, J., Wang, X. and Park, C.-S. (2016) 'A linked data system framework for sharing construction defect information using ontologies and BIM environments', Automation in Construction. Elsevier, 68, pp. 102–113. doi: 10.1016/J.AUTCON.2016.05.003.

Marsan, J., Paré, G. and Beaudry, A. (2012) 'Adoption of open source software in organizations: A socio-cognitive perspective', The Journal of Strategic Information Systems. North-Holland, 21(4), pp. 257–273. doi: 10.1016/J.JSIS.2012.05.004.

Misic, M. M. and Graf, D. K. (2004) 'Systems analyst activities and skills in the new millennium', Journal of Systems and Software, 71(1–2), pp. 31–36. doi: 10.1016/S0164-1212(02)00124-3.

Mohd Nawi, M. N., Baluch, N. and Bahauddin, A. Y. (2014) 'Impact of Fragmentation Issue in Construction Industry: An Overview', MATEC Web of Conferences, 15. doi: 10.1051/matecconf/20141501009.

Motik, B., Grau, B. C., Horrocks, I., Wu, Z., Fokoue, A. and Lutz, C. (2012) OWL 2 Web Ontology Language, W3C Recommendation. Available at: https://www.w3.org/TR/owl2-profiles/ (accessed 06/02/2018).

Musen, M. A. (2015) 'The Protégé project: A look back and a look forward', Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4). doi: 10.1145/2557001.25757003.

Mutis, I. and Issa, R. R. A. (2012) 'Framework for Semantic Reconciliation of Construction Project Information', Journal of Information Technology in Construction (ITcon), 17(2), pp. 1–24.

Nagy, D., Yassin, A. M. and Bhattacherjee, A. (2010) 'Organizational adoption of open source software: barriers and remedies', Communications of the ACM, 53(3), pp. 148–151. doi: 10.1145/1666420.1666457.

Niknam, M. and Karshenas, S. (2015) 'Integrating distributed sources of information for construction cost estimating using Semantic Web and Semantic Web Service technologies', Automation in Construction. Elsevier, 57, pp. 222–238. doi: 10.1016/J.AUTCON.2015.04.003.

O'Connor, M. J., Halaschek-Wiener, C. and Musen, M. A. (2010) 'Mapping Master: A Flexible Approach for Mapping Spreadsheets to OWL', in 9th International Semantic Web Conference. Shanghai.

Okoli, C. and Schabram, K. (2010) 'A Guide to Conducting a Systematic Literature Review of Information Systems Research', Working Papers on Information Systems, 10(26), pp.

1–51. doi: 10.2139/ssrn.1954824.

Omar, T. and Nehdi, M. L. (2016) ‘Data acquisition technologies for construction progress tracking’, Automation in Construction. Elsevier, 70, pp. 143–155. doi: 10.1016/J.AUTCON.2016.06.016.

Park, C.-S., Lee, D.-Y. and Kwon, O.-S. (2013) ‘A framework for proactive construction defect management using BIM, augmented reality and ontology-based data collection template’, Automation in Construction. Elsevier, 33, pp. 61–71. doi: 10.1016/J.AUTCON.2012.09.010.

Pauwels, P., Deursen, D. Van, Roo, J. De, Ackere, T. Van, Meyer, R. De, Walle, R. Van de and Campenhout, J. Van (2011) ‘Three-dimensional information exchange over the semantic web for the domain of architecture, engineering, and construction’, Representing and Reasoning About Three-Dimensional Space, 25(4), pp. 317–332.

Pauwels, P., Deursen, D. Van, Verstraeten, R., Roo, J. De, Meyer, R. De, Walle, R. Van de and Campenhout, J. Van (2011) ‘A semantic rule checking environment for building performance checking’, Automation in Construction, 20(5), pp. 506–518.

Pauwels, P., Zhang, S. and Lee, Y. C. (2017) ‘Semantic web technologies in AEC industry: A literature overview’, Automation in Construction. Elsevier B.V., 73, pp. 145–165. doi: 10.1016/j.autcon.2016.10.003.

Qady, M. Al and Kandil, A. (2012) ‘Document Discourse for Managing Construction Project Documents’, Journal of Computing in Civil Engineering, 27(5), pp. 466–475.

Rezgui, Y., Boddy, S., Wetherill, M. and Cooper, G. (2011) ‘Past, present and future of information and knowledge sharing in the construction industry: Towards semantic service-based e-construction?', Computer-Aided Design. Elsevier, 43(5), pp. 502–515. doi: 10.1016/J.CAD.2009.06.005.

Rietveld, L. and Hoekstra, R. (2014) 'YASGUI: Feeling the Pulse of Linked Data', in Janowicz, K., Schlobach, S., Lambrix, P., and Hyvönen, E. (eds) Knowledge Engineering and Knowledge Management. EKAW 2014. Linköping: Springer, Cham, pp. 441–452. doi: https://doi.org/10.1007/978-3-319-13704-9_34.

Roman, D., Nikolov, N., Putlier, A., Sukhobok, D., Elvesæter, B., Berre, A., Ye, X., Dimitrov, M., Simov, A., Zarev, M., Moynihan, R., Roberts, B., Berlocher, I., Kim, S., Lee, T., Smith, A. and Heath, T. (2017) 'DataGraft: One-stop-shop for open data management', Semantic Web, pp. 1–19. doi: 10.3233/SW-170263.

Rubin, V. L. (2014) 'Veracity roadmap: Is big data objective, truthful and credible?', Advances in Classification Research Online, 24, pp. 4–15. doi: 10.7152/acro.v24i1.14671.

Rusu, O., Halcu, I., Grigoriu, O., Neculoiu, G., Sandulescu, V., Marinescu, M. and Marinescu, V. (2013) 'Converting unstructured and semi-structured data into knowledge', in 11th RoEduNet International Conference. Sinaia, pp. 1–4. doi: 10.1109/RoEduNet.2013.6511736.

Sarrab, M. and Rehman, O. M. H. (2014) 'Empirical study of open source software selection for adoption, based on software quality characteristics', Advances in Engineering Software. Elsevier, 69, pp. 1–11. doi: 10.1016/J.ADVENGSOFT.2013.12.001.

Shayeganfar, F., Mahdavi, A., Suter, G. and Anjomshoaa, A. (2009) 'Implementation of an IFD library using semantic web technologies: A case study', Ework and Ebusiness in Architecture, Engineering and Construction, pp. 539–544.

Soibelman, L., Wu, J., Caldas, C., Brilakis, I. and Lin, K.-Y. (2008) 'Management and analysis of unstructured construction data types', Advanced Engineering Informatics. Elsevier, 22(1), pp. 15–27. doi: 10.1016/J.AEI.2007.08.011.

Soleimanifar, M. (2016) Integrated Project Management Framework for a Pipe Spool Fabrication Shop. University of Alberta. doi: doi:10.7939/R36M33G31.

Son, H., Bosché, F. and Kim, C. (2015) 'As-built data acquisition and its use in production monitoring and automated layout of civil infrastructure: A survey', Advanced Engineering Informatics. Elsevier Ltd, 29(2), pp. 172–183. doi: 10.1016/j.aei.2015.01.009.

Standards Norway (1996) Piping Fabrication, Installation, Flushing and Testing. NORSOK. Available at: http://www.ccc.no/publish_files/Norsok_L-004-CR_Ror_og_trykk_Engelsk.pdf (accessed 06/02/2018).

Sun, M. and Aouad, G. (1999) 'Control Mechanism for Information Sharing in an Integrated Construction Environment', in Proceeding of The 2nd International Conference on Concurrent Engineering in Construction, pp. 1–10.

TechInvestLab.ru (2013) .15926 Editor. Available at: http://techinvestlab.ru/files/V4/dot15926 Editor14_Vol4_PatternsAndMapping.pdf (accessed 13/02/2018).

Tennison, J., Open Data Institute, Kellogg, G., Kellogg Associates, Herman, I. and W3C (2015) Model for Tabular Data and Metadata on the Web, W3C Recommendation. Available at:

http://www.w3.org/TR/tabular-data-model/ (accessed 14/02/2018).

The W3C SPARQL Working Group (2013) SPARQL 1.1 Overview, W3C Recommendation. Available at: https://www.w3.org/TR/sparql11-overview/ (accessed 06/02/2018).

Walsh, C. (2016) Data and Analytics: Open Source Data Integration Tool Comparison. Available at: https://www.excella.com/wp-content/uploads/2016/03/Open-Source-DI-Tool-Comparison_March2016.pdf (accessed 06/02/2018).

Westin, S. and Sein, M. K. (2014) 'Improving Data Quality in Construction Engineering Projects: An Action Design Research Approach', Journal of Management in Engineering, 30(3), pp. 1–11. doi: 10.1061/(ASCE)ME.1943-5479.0000202.

Wiesner, A., Morbach, J. and Marquardt, W. (2011) 'Information integration in chemical process engineering based on semantic technologies', Computers and Chemical Engineering. Elsevier Ltd, 35(4), pp. 692–708. doi: 10.1016/j.compchemeng.2010.12.003.

Yang, R. (2009) Process Plant Lifecycle Information Management. Bloomington, IN: iUniverse.

Yurchyshyna, A. and Zarli, A. (2009) 'An ontology-based approach for formalisation and semantic organisation of conformance requirements in construction', Automation in Construction. Elsevier B.V., 18(8), pp. 1084–1098. doi: 10.1016/j.autcon.2009.07.008.

Zhai, D., Goodrum, P. M., Haas, C. T. and Caldas, C. H. (2009) 'Relationship between Automation and Integration of Construction Information Systems and Labor

Productivity', Journal of Construction Engineering and Management, 135(8).

Zhang, J., Li, H., Zhao, Y. and Ren, G. (2018) 'An ontology-based approach supporting holistic structural design with the consideration of safety, environmental impact and cost', Advances in Engineering Software. Elsevier, 115, pp. 26–39. doi: 10.1016/J.ADVENGSOFT.2017.08.010.

Zhang, L. and Issa, R. (2011) 'Development of IFC-based construction industry ontology for information retrieval from IFC Models', in Proceedings of the 2011 eg-ice Workshop, University of Twente, The Netherlands, July 6–8.

Zhong, B. T., Ding, L. Y., Love, P. E. D. and Luo, H. B. (2015) 'An ontological approach for technical plan definition and verification in construction', Automation in Construction, 55, pp. 47–57. doi: 10.1016/j.autcon.2015.02.002.

**Table 1.** Definition of key terms

| Term | Definition | Reference |
|---|---|---|
| Big Data | large and complex data that cannot be managed by traditional data tools | (Bilal, Oyedele, Akinade, et al., 2016) |
| CSV | computer file format for storing tabular data in plain text | (Tennison et al., 2015) |
| Database | a structured set of data held in a computer | (Doan et al., 2012) |
| Data Model | the relationship structure of the entities in a data source | (Doan et al., 2012) |
| EDMS | software for creation, storage and control of documents electronically | (Qady and Kandil, 2012) |
| EMS | software that supports business processes in complex organisations | (Sun and Aouad, 1999) |
| Ontology | description of the specific contents of a given domain | (Curé and Blin, 2015) |
| Open Source | freely available | (Marsan et al., 2012) |
| Relational Database | a database type where structured data is stored in related tables | (Doan et al., 2012) |
| RDF | metadata (data about data) used for describing information resources | (Curé and Blin, 2015) |
| Semantic Web | an extension of the World Wide Web that enhances information retrieval | (Curé and Blin, 2015) |
| Structured Data | information with a high degree of organisation | (Doan et al., 2012) |
| Systems Analyst | a person who uses information technology to improve efficiency of complex processes | (Misic and Graf, 2004) |
| XML | annotation language for storing and transmitting data | (Fiatech, 2011) |

**Table 2.** Ontology-based data integration applications in construction process

| Construction Process | Activity | References |
|---|---|---|
| Planning | Technical planning | (Zhong et al., 2015) |
| | Sustainable construction planning | (Zhang et al., 2018) |
| | Risk planning | (Ding et al., 2016) |
| | Cost estimation | (Abanda et al., 2011) |
| Compliance | Code Checking | (Zhang and Issa, 2011; Yurchyshyna and Zarli, 2009) |
| Monitoring | Performance checking | (Pauwels, Deursen, Verstraeten et al., 2011) |
| Control | Defect management | (Park et al., 2013) |
| General | 3D model integration | (Pauwels, Deursen, Roo et al., 2011) |
| | Project information management | (Elghamrawy et al., 2009) |

**Figure 1.** Research design

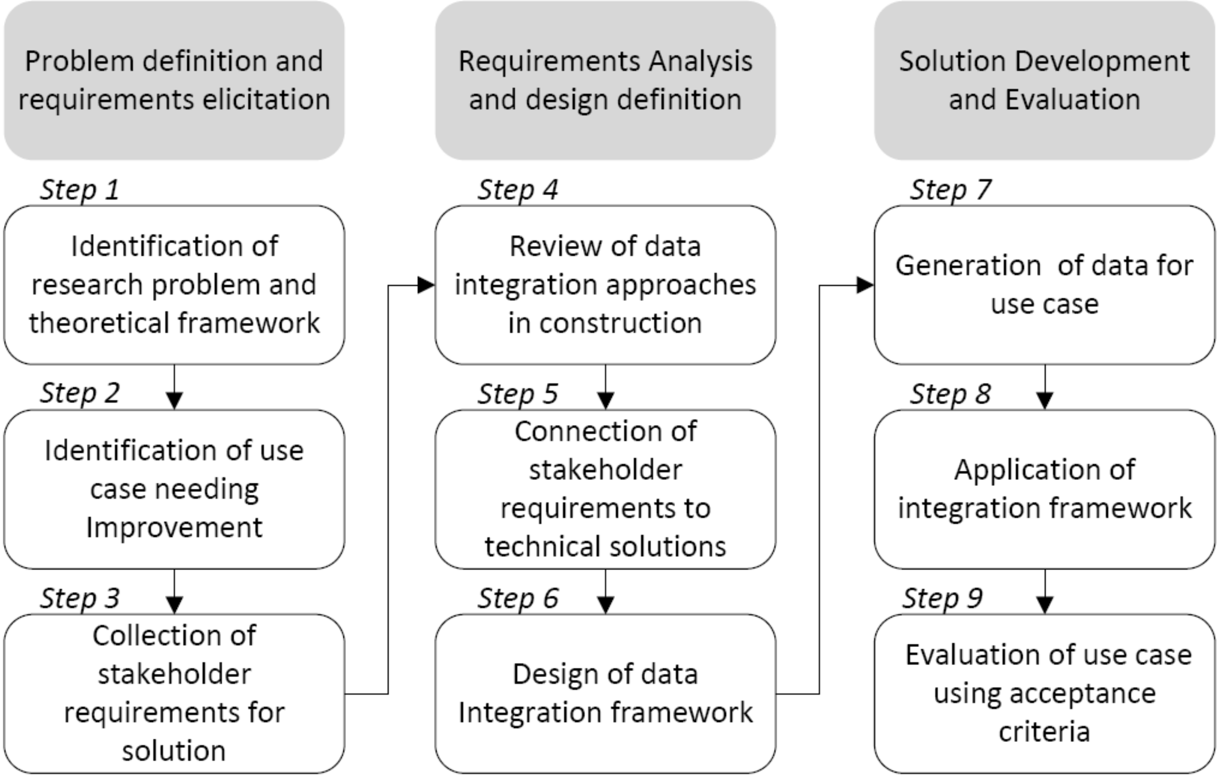

Problem definition and requirements elicitation
Requirements Analysis and design definition
Solution Development and Evaluation
Step 1
Identification of research problem and theoretical framework
Step 2
Identification of use case needing Improvement
Step 3
Collection of stakeholder requirements for solution
Step 4
Review of data integration approaches in construction
Step 5
Connection of stakeholder requirements to technical solutions
Step 6
Design of data Integration framework
Step 7
Generation of data for use case
Step 8
Application of integration framework
Step 9
Evaluation of use case using acceptance criteria

**Figure 2.** Extracting value from construction data with big data technology

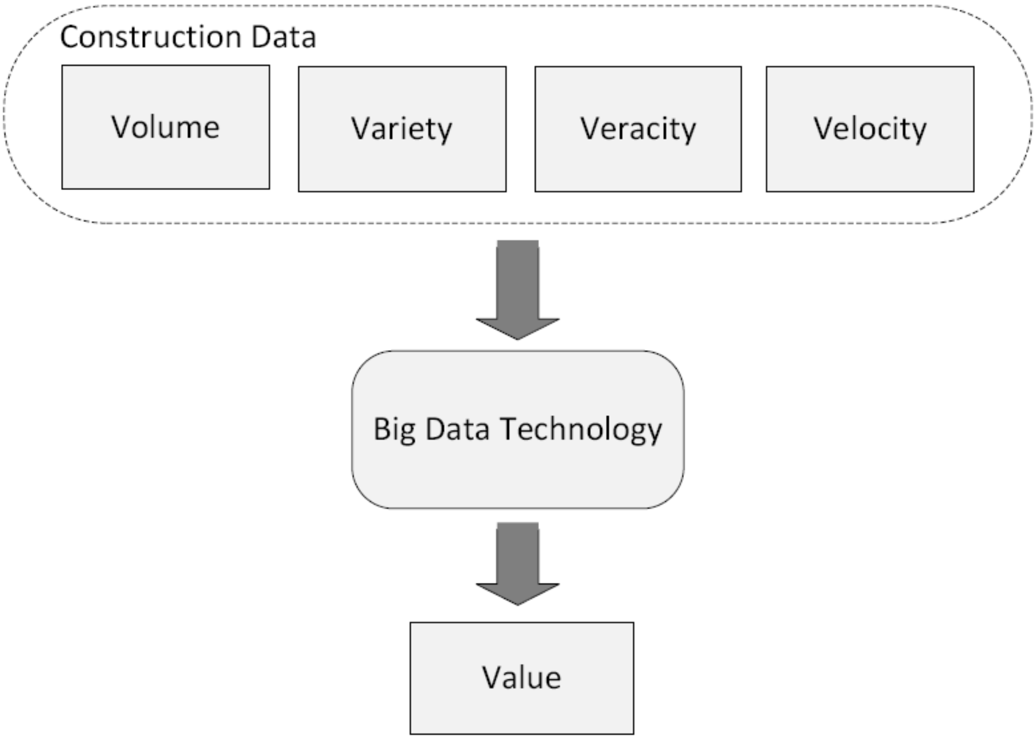


**Figure 3.** Illustration of construction data integration techniques

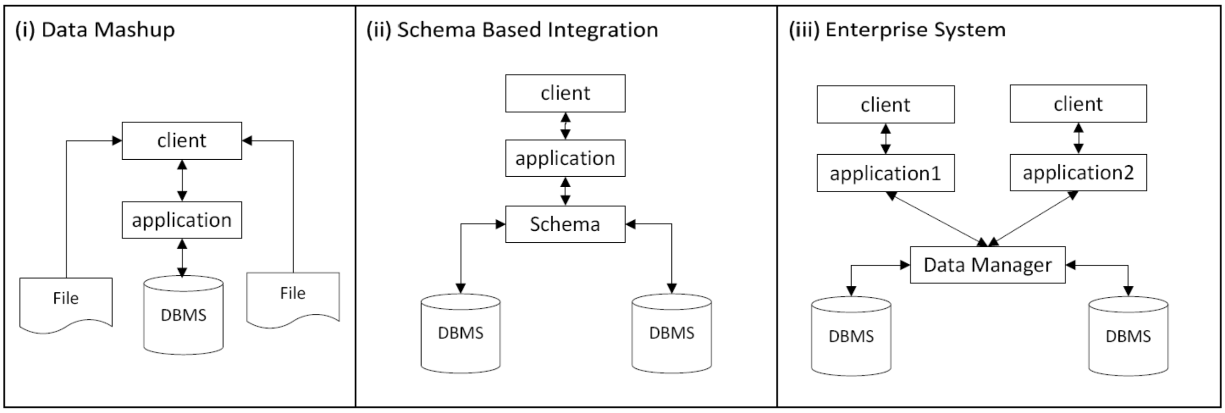


**Figure 4.** Integration framework for ad hoc data projects during construction

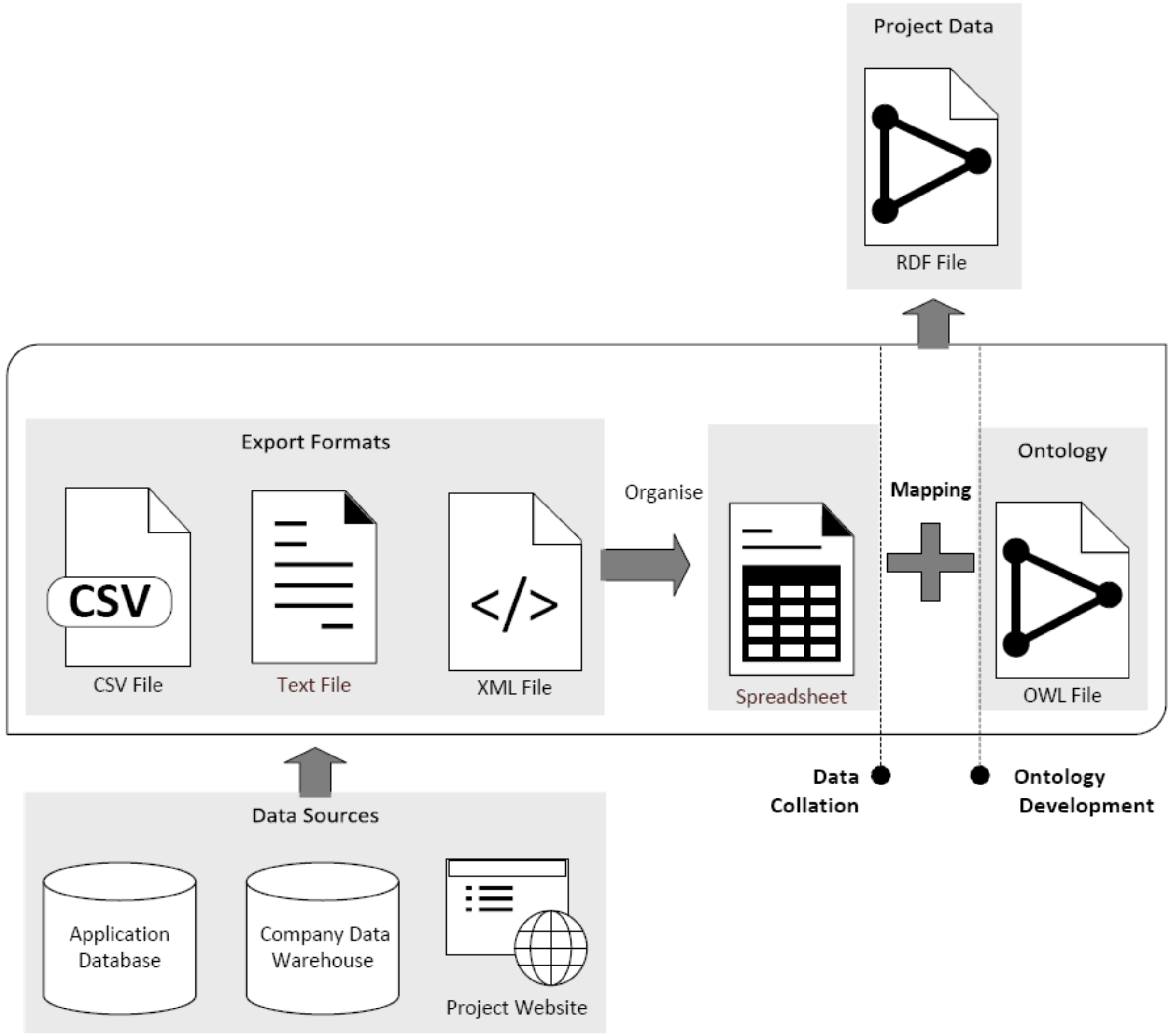
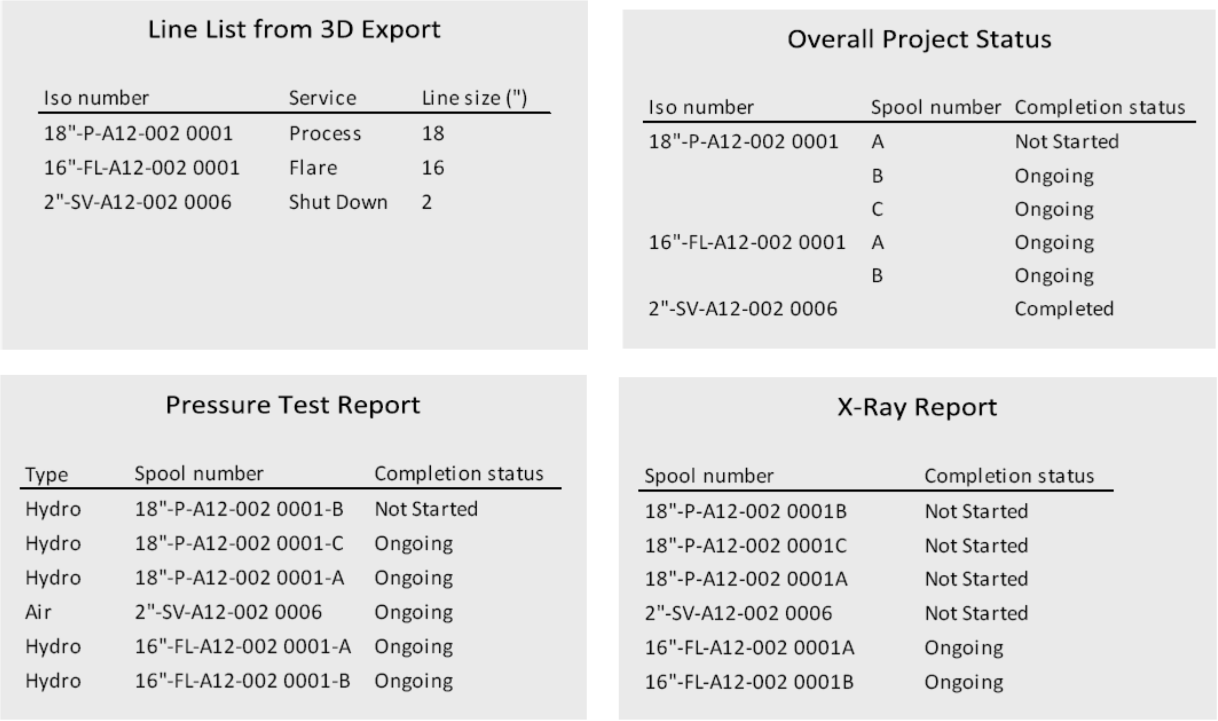


**Figure 5.** Sample data about 6 spools from different construction project sources



Line List from 3D Export

| Iso number | Service | Line size (") |
|---|---|---|
| 18"-P-A12-002 0001 | Process | 18 |
| 16"-FL-A12-002 0001 | Flare | 16 |
| 2"-SV-A12-002 0006 | Shut Down | 2 |

Overall Project Status

| Iso number | Spool number | Completion status |
|---|---|---|
| 18"-P-A12-002 0001 | A | Not Started |
|  | B | Ongoing |
|  | C | Ongoing |
| 16"-FL-A12-002 0001 | A | Ongoing |
|  | B | Ongoing |
| 2"-SV-A12-002 0006 |  | Completed |

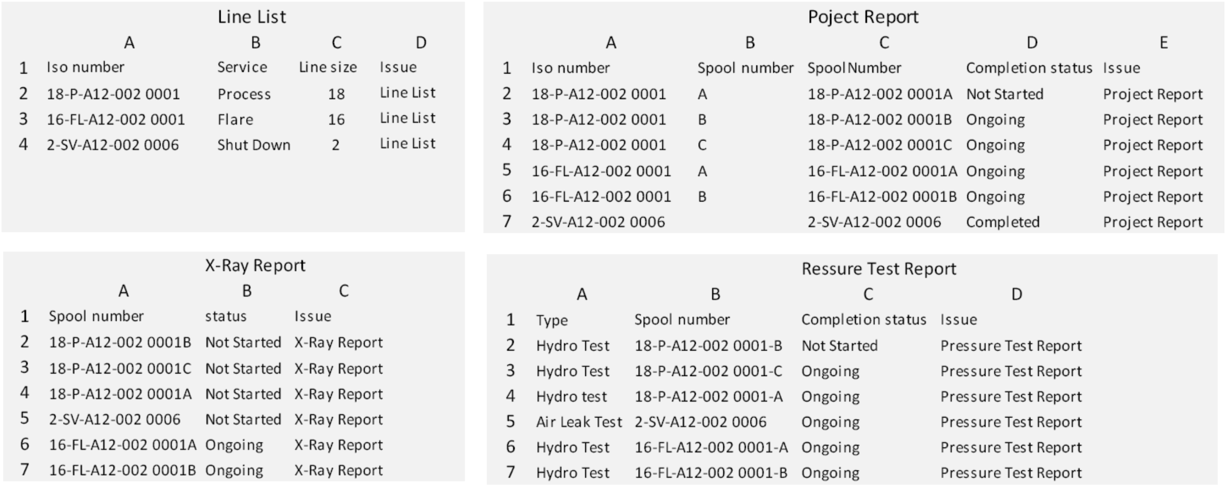
Pressure Test Report

| Type | Spool number | Completion status |
|---|---|---|
| Hydro | 18"-P-A12-002 0001-B | Not Started |
| Hydro | 18"-P-A12-002 0001-C | Ongoing |
| Hydro | 18"-P-A12-002 0001-A | Ongoing |
| Air | 2"-SV-A12-002 0006 | Ongoing |
| Hydro | 16"-FL-A12-002 0001-A | Ongoing |
| Hydro | 16"-FL-A12-002 0001-B | Ongoing |

X-Ray Report

| Spool number | Completion status |
|---|---|
| 18"-P-A12-002 0001B | Not Started |
| 18"-P-A12-002 0001C | Not Started |
| 18"-P-A12-002 0001A | Not Started |
| 2"-SV-A12-002 0006 | Not Started |
| 16"-FL-A12-002 0001A | Ongoing |
| 16"-FL-A12-002 0001B | Ongoing |

Figure 6. Processed data in spreadsheets



Line List

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Iso number | Service | Line size | Issue |
| 2 | 18-P-A12-002 0001 | Process | 18 | Line List |
| 3 | 16-FL-A12-002 0001 | Flare | 16 | Line List |
| 4 | 2-SV-A12-002 0006 | Shut Down | 2 | Line List |

Poject Report

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Iso number | Spool number | SpoolNumber | Completion status | Issue |
| 2 | 18-P-A12-002 0001 | A | 18-P-A12-002 0001A | Not Started | Project Report |
| 3 | 18-P-A12-002 0001 | B | 18-P-A12-002 0001B | Ongoing | Project Report |
| 4 | 18-P-A12-002 0001 | C | 18-P-A12-002 0001C | Ongoing | Project Report |
| 5 | 16-FL-A12-002 0001 | A | 16-FL-A12-002 0001A | Ongoing | Project Report |
| 6 | 16-FL-A12-002 0001 | B | 16-FL-A12-002 0001B | Ongoing | Project Report |
| 7 | 2-SV-A12-002 0006 | | 2-SV-A12-002 0006 | Completed | Project Report |

X-Ray Report

| | A | B | C |
|---|---|---|---|
| 1 | Spool number | status | Issue |
| 2 | 18-P-A12-002 0001B | Not Started | X-Ray Report |
| 3 | 18-P-A12-002 0001C | Not Started | X-Ray Report |
| 4 | 18-P-A12-002 0001A | Not Started | X-Ray Report |
| 5 | 2-SV-A12-002 0006 | Not Started | X-Ray Report |
| 6 | 16-FL-A12-002 0001A | Ongoing | X-Ray Report |
| 7 | 16-FL-A12-002 0001B | Ongoing | X-Ray Report |

Ressure Test Report

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Type | Spool number | Completion status | Issue |
| 2 | Hydro Test | 18-P-A12-002 0001-B | Not Started | Pressure Test Report |
| 3 | Hydro Test | 18-P-A12-002 0001-C | Ongoing | Pressure Test Report |
| 4 | Hydro test | 18-P-A12-002 0001-A | Ongoing | Pressure Test Report |
| 5 | Air Leak Test | 2-SV-A12-002 0006 | Ongoing | Pressure Test Report |
| 6 | Hydro Test | 16-FL-A12-002 0001-A | Ongoing | Pressure Test Report |
| 7 | Hydro Test | 16-FL-A12-002 0001-B | Ongoing | Pressure Test Report |

**Figure 7.** Spool ontology

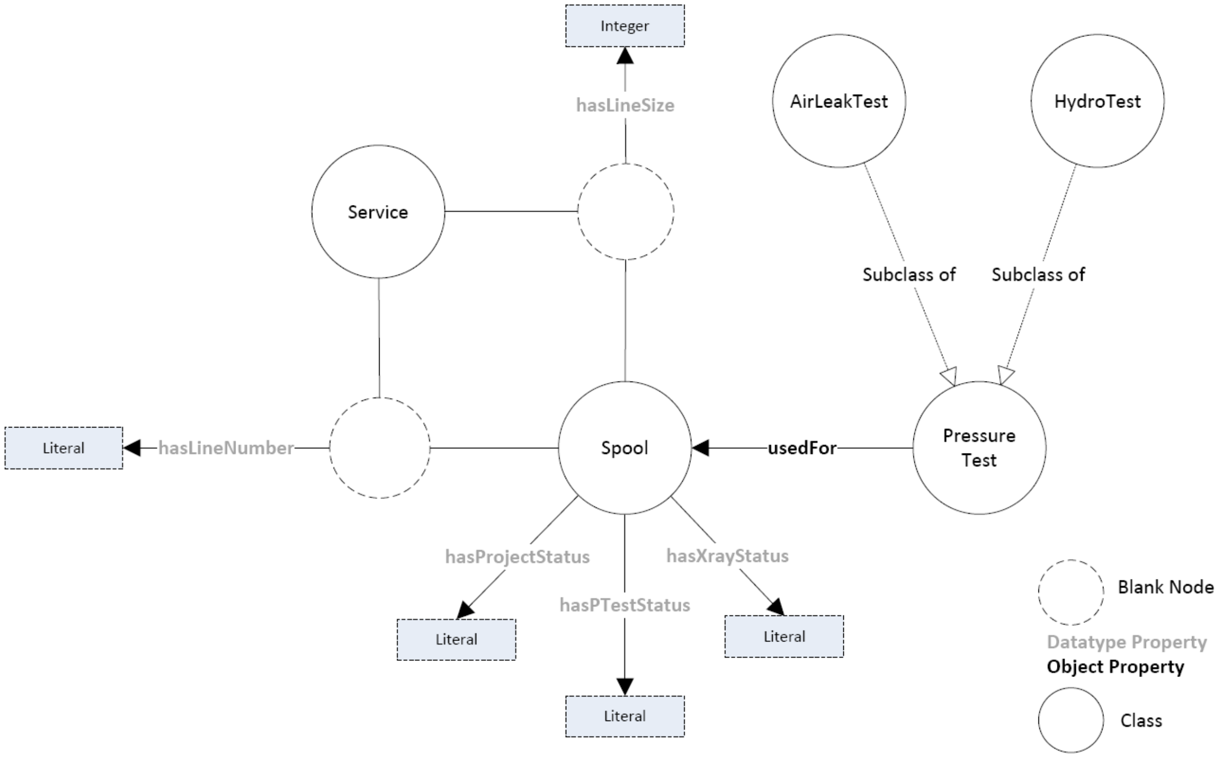
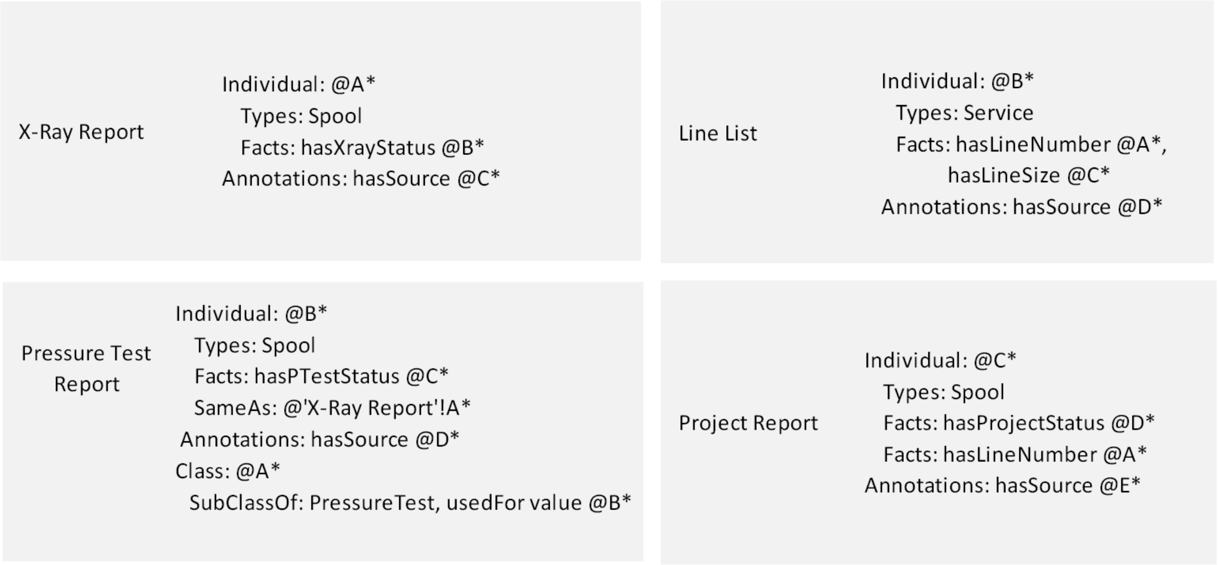


**Figure 8.** Data transformation rules for mappings to ontology



X-Ray Report

```
Individual: @A*
   Types: Spool
   Facts: hasXrayStatus @B*
Annotations: hasSource @C*
```

Line List

```
Individual: @B*
   Types: Service
   Facts: hasLineNumber @A*,
          hasLineSize @C*
Annotations: hasSource @D*
```

Pressure Test Report

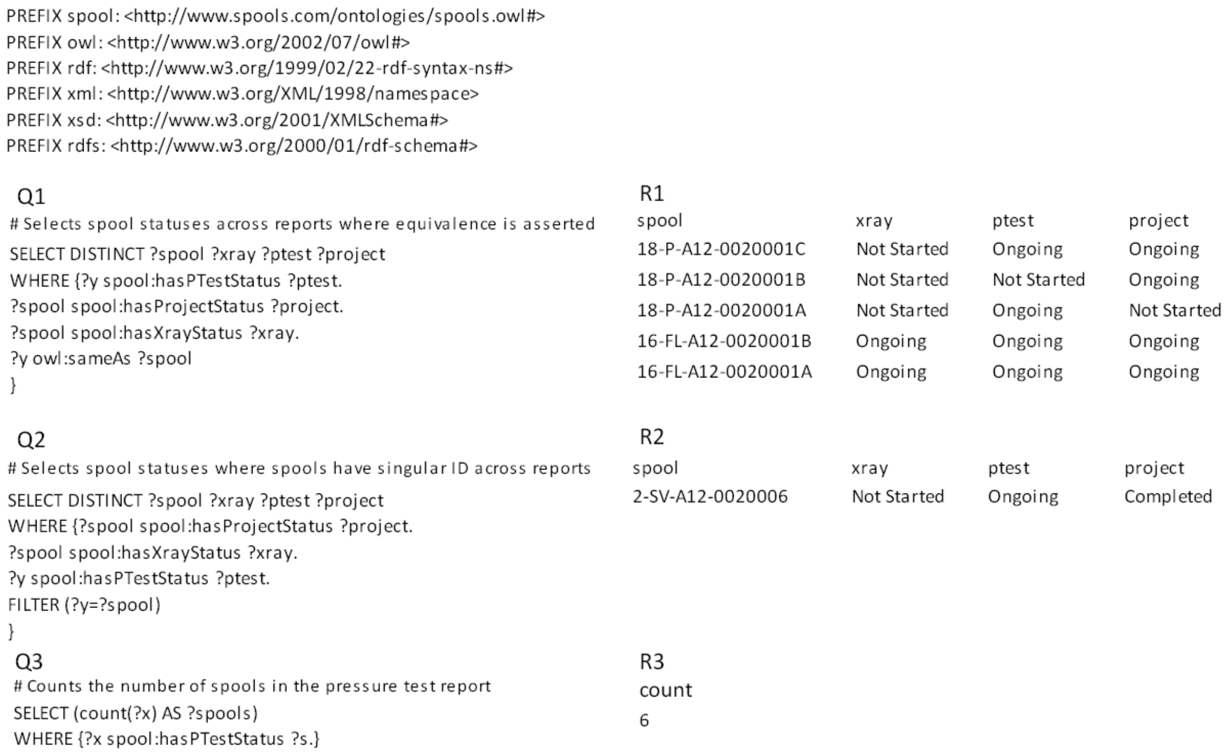
```
Individual: @B*
   Types: Spool
   Facts: hasPTestStatus @C*
   SameAs: @'X-Ray Report'!A*
 Annotations: hasSource @D*
Class: @A*
   SubClassOf: PressureTest, usedFor value @B*
```

Project Report

```
Individual: @C*
   Types: Spool
   Facts: hasProjectStatus @D*
   Facts: hasLineNumber @A*
Annotations: hasSource @E*
```

* refers to all data in referenced spreadsheet column

**Figure 9.** Sample queries on integrated data



```
PREFIX spool: <http://www.spools.com/ontologies/spools.owl#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xml: <http://www.w3.org/XML/1998/namespace>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```

Q1

```
# Selects spool statuses across reports where equivalence is asserted
SELECT DISTINCT ?spool ?xray ?ptest ?project
WHERE {?y spool:hasPTestStatus ?ptest.
?spool spool:hasProjectStatus ?project.
?spool spool:hasXrayStatus ?xray.
?y owl:sameAs ?spool
}
```

R1

| spool | xray | ptest | project |
|---|---|---|---|
| 18-P-A12-0020001C | Not Started | Ongoing | Ongoing |
| 18-P-A12-0020001B | Not Started | Not Started | Ongoing |
| 18-P-A12-0020001A | Not Started | Ongoing | Not Started |
| 16-FL-A12-0020001B | Ongoing | Ongoing | Ongoing |
| 16-FL-A12-0020001A | Ongoing | Ongoing | Ongoing |

Q2

```
# Selects spool statuses where spools have singular ID across reports
SELECT DISTINCT ?spool ?xray ?ptest ?project
WHERE {?spool spool:hasProjectStatus ?project.
?spool spool:hasXrayStatus ?xray.
?y spool:hasPTestStatus ?ptest.
FILTER (?y=?spool)
}
```

R2

| spool | xray | ptest | project |
|---|---|---|---|
| 2-SV-A12-0020006 | Not Started | Ongoing | Completed |

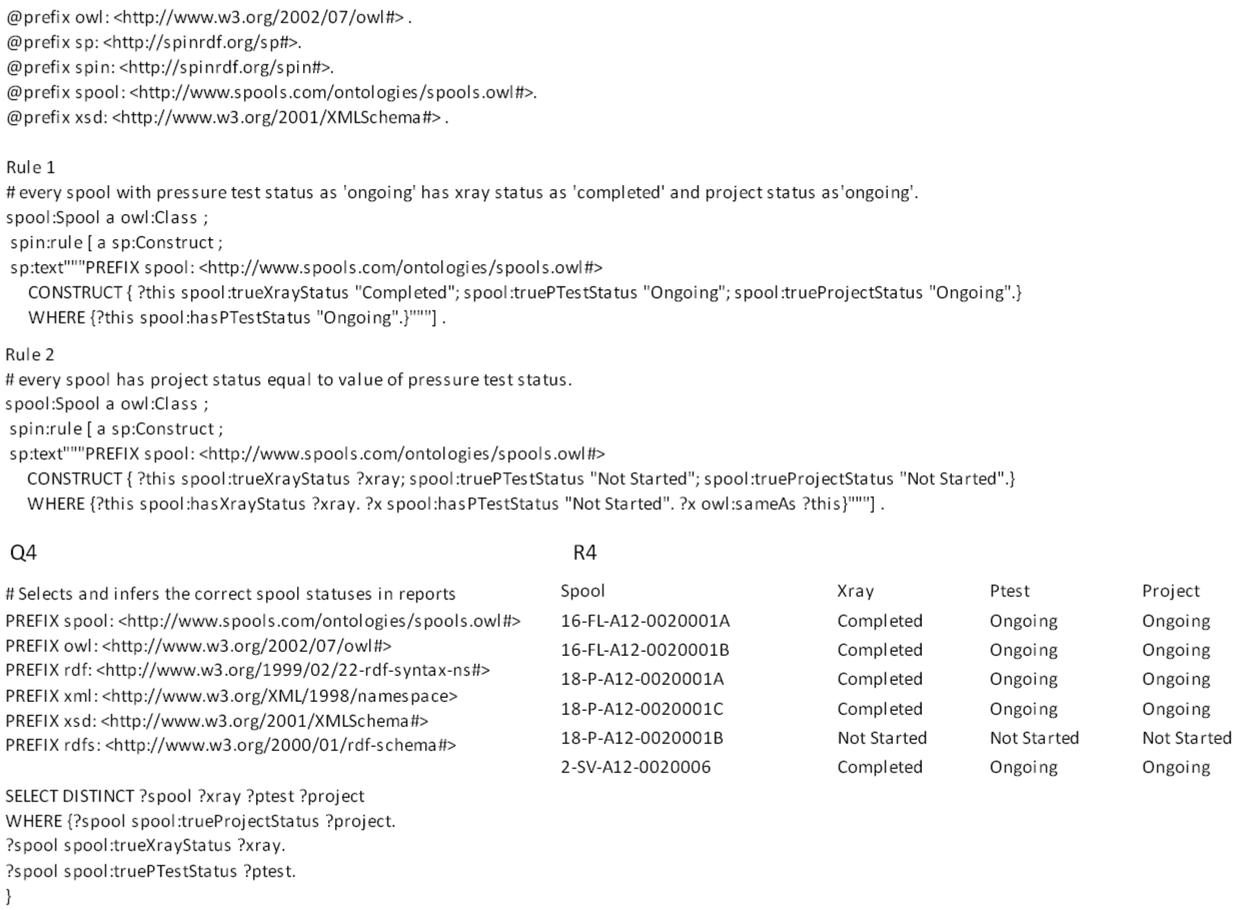
Q3

```
# Counts the number of spools in the pressure test report
SELECT (count(?x) AS ?spools)
WHERE {?x spool:hasPTestStatus ?s.}
```

R3

| count |
|---|
| 6 |

**Figure 10.** Rules and sample query with inference on integrated data to detect data quality issues



```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix sp: <http://spinrdf.org/sp#>.
@prefix spin: <http://spinrdf.org/spin#>.
@prefix spool: <http://www.spools.com/ontologies/spools.owl#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

Rule 1
# every spool with pressure test status as 'ongoing' has xray status as 'completed' and project status as'ongoing'.
spool:Spool a owl:Class ;
 spin:rule [ a sp:Construct ;
 sp:text"""PREFIX spool: <http://www.spools.com/ontologies/spools.owl#>
    CONSTRUCT { ?this spool:trueXrayStatus "Completed"; spool:truePTestStatus "Ongoing"; spool:trueProjectStatus "Ongoing".}
    WHERE {?this spool:hasPTestStatus "Ongoing".}"""] .

Rule 2
# every spool has project status equal to value of pressure test status.
spool:Spool a owl:Class ;
 spin:rule [ a sp:Construct ;
 sp:text"""PREFIX spool: <http://www.spools.com/ontologies/spools.owl#>
    CONSTRUCT { ?this spool:trueXrayStatus ?xray; spool:truePTestStatus "Not Started"; spool:trueProjectStatus "Not Started".}
    WHERE {?this spool:hasXrayStatus ?xray. ?x spool:hasPTestStatus "Not Started". ?x owl:sameAs ?this}"""] .
```

Q4

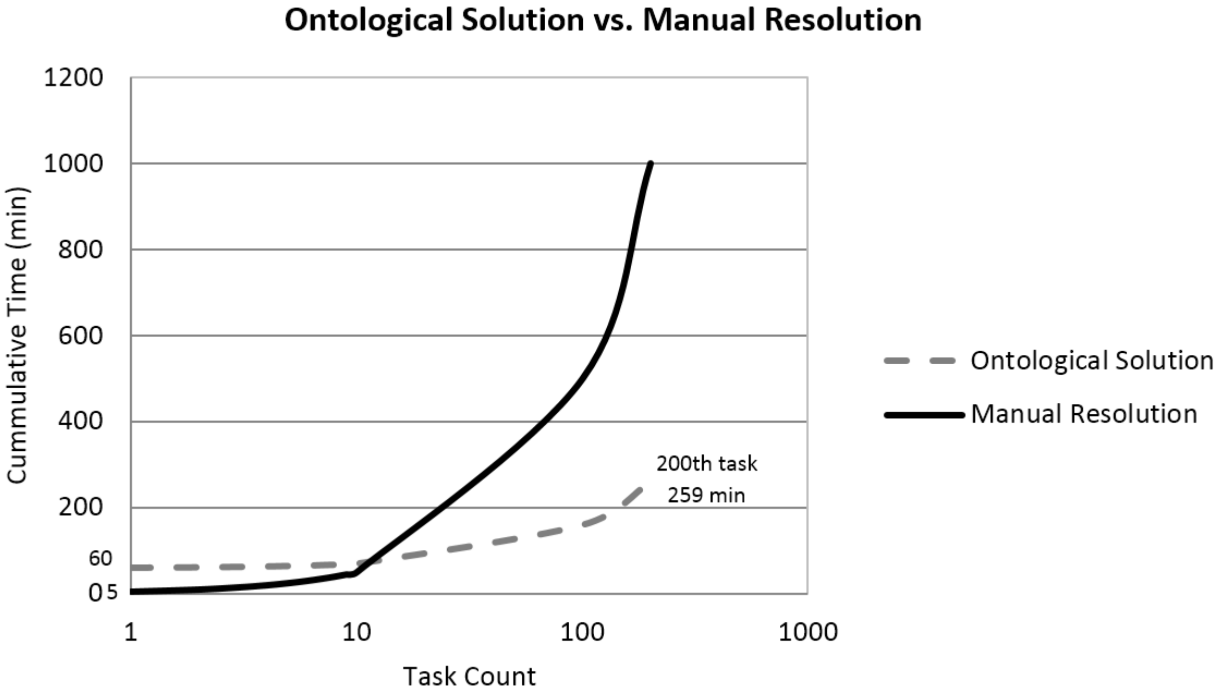
```
# Selects and infers the correct spool statuses in reports
PREFIX spool: <http://www.spools.com/ontologies/spools.owl#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xml: <http://www.w3.org/XML/1998/namespace>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?spool ?xray ?ptest ?project
WHERE {?spool spool:trueProjectStatus ?project.
?spool spool:trueXrayStatus ?xray.
?spool spool:truePTestStatus ?ptest.
}
```

R4

| Spool | Xray | Ptest | Project |
|---|---|---|---|
| 16-FL-A12-0020001A | Completed | Ongoing | Ongoing |
| 16-FL-A12-0020001B | Completed | Ongoing | Ongoing |
| 18-P-A12-0020001A | Completed | Ongoing | Ongoing |
| 18-P-A12-0020001C | Completed | Ongoing | Ongoing |
| 18-P-A12-0020001B | Not Started | Not Started | Not Started |
| 2-SV-A12-0020006 | Completed | Ongoing | Ongoing |

**Figure 11.** A comparison of task cumulative duration between ontological solution and manual process


Ontological Solution vs. Manual Resolution
Cummulative Time (min)
Task Count
200th task
259 min
Ontological Solution
Manual Resolution