PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Universidade Federal do Rio de Janeiro

Kelli de Faria Cordeiro

aDApTA: ADAPTIVE APPROACH FOR INFORMATION INTEGRATION TO SUPPORT DECISION MAKING IN COMPLEX ENVIRONMENTS

Doctoral Thesis

Instituto de Matemática

Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
INSTITUTO TÉRCIO PACITTI DE APLICAÇÕES E PESQUISAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

KELLI DE FARIA CORDEIRO

aDApTA: ADAPTIVE APPROACH FOR INFORMATION INTEGRATION TO SUPPORT DECISION MAKING IN COMPLEX ENVIRONMENTS

Doctoral's thesis submitted to the Programa de Pós-Graduação em Informática, Instituto de Matemática, Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Universidade Federal do Rio de Janeiro as a partial requirement to obtain the title of Doctor in Informatics.

Advisors: Profa. Maria Luiza Machado Campos, Ph.D

Prof. Marcos Roberto da Silva Borges, Ph.D

Rio de Janeiro
2015

CIP - Catalogação na Publicação

C794a

Cordeiro, Kelli de Faria
aDApTA: Adaptive Approach for Information Integration to Support Decision Making in Complex Environments / Kelli de Faria Cordeiro. -- Rio de Janeiro, 2015.
148 f.

Orientadora: Maria Luiza Machado Campos.
Coorientador: Marcos Roberto da Silva Borges.
Tese (doutorado) - Universidade Federal do Rio de Janeiro, Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Programa de Pós Graduação em informática, 2015.

1. Information Integration. 2. Complex Information System. 3. Adaptation. 4. Decision Support System. 5. Linked Data. I. Campos, Maria Luiza Machado, orient. II. Borges, Marcos Roberto da Silva, coorient. III. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a).

aDApTA: Adaptive Approach for Information Integration to Support Decision Making in Complex Environment

Kelli de Faria Cordeiro

Tese de Doutorado submetida ao Programa de Pós-Graduação em Informática do Instituto de Matemática e do Instituto Tércio Pacciti da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários à obtenção do título de Doutora em Informática.

Aprovada em 24/03/2015 por:

Prof. Marcos Roberto da Silva Borges, Ph.D., UFRJ (Presidente)

Profa. Maria Luiza Machado Campos, Ph.D., UFRJ

Profa. Jonice de Oliveira Sampaio, D.Sc., UFRJ

Profa. Adriana Santarosa Vivacqua, D.Sc., UFRJ

Prof. Jano Moreira de Souza, Ph.D., UFRJ

Profa. Maria Claudia Reis Cavalcanti, D.Sc., IME

Profa. Ana Carolina Brandão Salgado, Docteur, UFPE

Dedico à Deus
Sólido alicerce
Vida segura

Acknowledgments

Doutorado é um projeto que conta com a contribuição de muitas pessoas, formando uma grande equipe. Sempre soube disso, mas nunca imaginei que a equipe fosse tão numerosa e que a participação dos seus membros fosse tão decisiva! O produto final desta longa jornada e desse desafiador empreendimento tem a participação de todos os senhores e senhoras, familiares, amigos e chefes.

Creio que o início de tudo está na nossa base familiar, vejo o reflexo dos meus pais presentes no processo de desenvolvimento deste trabalho. A inquietação, curiosidade, determinação e perseverança da minha mãe, Maria Gloria de Faria, e a disciplina, comprometimento, humildade e simplicidade do meu pai, Geraldo Cordeiro. Nessa base, estão as minhas irmãs, Renata e Hebe, que sempre demonstraram admiração e disponibilidade para qualquer possível contribuição. O apoio incondicional diante de tantas adversidades e a compreensão pela ausência foram decisivos!

Mesmo diante de tantas ausências, sempre pude contar com meus amigos e amigas, que são família, para aquela saída para relaxar, rir e descontrair. Juliana, Adriana, Bárbara, Hugo, Barbosa, Keila, Thaynara, Luzimar, Thomas, e, em especial, a Kilza, a que mais aguentou tantos altos e baixos e não desistiu de mim. Agradeço também à Cida por cuidar dos meus afazeres domésticos. O apoio, carinho, compreensão e presença de vocês foram decisivos!

A maior parte do período do desenvolvimento deste trabalho foi conciliada com as minhas funções e responsabilidades na Marinha do Brasil. Na verdade, este trabalho nem teria começado se não fosse o

incentivo e apoio do Almirante Edesio, Comandantes Marisa, Hamilton e Maria Angélica. E a continuidade e a conclusão só foram possíveis com o empenho pessoal dos Almirantes Jayme e Hugo, Comandantes Azevedo, Lage, Raquel, e Funcionárias Civis Lúcia e Ivana. Agradeço também a tranquilidade que meu orientador técnico, Comandante Muradas, me passou para a conclusão deste trabalho. As decisões e a confiança dos senhores e senhoras foram vitais. Espero poder honrá-las!

Ainda na Marinha, pude contar com amigos e amigas, que também são família, para as comemorações diante de cada etapa alcançada e principalmente para o apoio nos momentos críticos e de grande pressão. André Vitor, Beth, Camila, Virginia, Christina, João Luis, Denise, Viviane Celso, Maurício, Manoel Ribeiro, Lucimar, Rufino e Rodrigo. O papel de vocês nessa jornada e na minha vida é essencial!

Especial agradecimento aos primeiros membros desta equipe, meus orientadores, Professores Marcos Borges e Maria Luiza, incluindo meu orientador de Mestrado, Professor Pedro Manoel, e minha amiga Glenda, quem me apresentou o programa.

A relação orientador-orientado é muito próxima, sempre soube isso, mas não tinha ideia que em um Doutorado seria tanto! É uma relação de confiança, respeito e admiração. Obrigada Professores pela enorme compreensão diante das adversidades ao longo do caminho. A maneira cordial como sempre me trataram, os incentivos e os desafios apresentados foram determinantes para concluir este trabalho e para a minha formação como pesquisadora!

Ainda na Universidade, também pude contar com muitos amigos para o desenvolvimento das pesquisas e descoberta de muito conhecimento. João Moreira, Inês Boscá, Fabrício Firmino, Veruska, Miguel Grabiel, Alan Tygel, Tiago Marino, Bruna Dürr, e vários outros, obrigada por todas as oportunidades de boas discussões. Nas implementações, pude contar com o excelente trabalho e apoio do amigo Rogers, alunos de Iniciação Científica, Camila e João, e o técnico Thiago Ferreira. Contei também com a presteza e competência do Anibal na resolução de variadas questões administrativos. Obrigada toda equipe do programa que fizeram esse curso acontecer!

Por fim, agradeço aos Professores(as) Jano, Ana Carolina, Yoko, Jonice, Adriana pela disponibilidade para participação da banca de avaliação dessa tese, e também aos membros suplentes, Paulo Vitor e Flavia Santoro. As críticas dos senhores serão vitais para o aprimoramento desta pesquisa.

Obrigada notebook por ter aguentado o tranco!

Obrigada, Deus!

Special Acknowledgments

Preciso de um capítulo a parte para agradecer a primeira pessoa que, generosamente, me acolheu no programa e, de forma muito sábia, me conduziu até aqui, a Profa Maria Luiza. Mesmo sendo um capítulo a parte, creio que o agradecimento será insuficiente e incompleto diante da complexidade dos seus ensinamentos, que fazem uso de um dos métodos mais eficazes de aprendizagem, o exemplo.

Como aluna, tive a oportunidade de observar seu trabalho apaixonado de motivar, provocar e ensinar, seu empenho e perseverança com aprendizado de cada aluno, seu cuidado e sensibilidade com o humano por trás dos momentos de dificuldade, suas visões e ideologias acadêmicas, seu processo sistemático e empolgado de pesquisar, seu zelo com cada produto desenvolvido, sua realização com cada conquista alcançada pelos alunos, seu cuidado e sutileza com cada crítica, sua exigência pelo aprimoramento de detalhes dos trabalhos. E em todas as ocasiões, o bom humor! Como este trabalho foi desenvolvido com vários momentos de boas gargalhadas.

Também tive a oportunidade de compartilhar do seu convívio familiar. Obrigada pelo acolhimento Arnaldo, Júlia e Luiza durante os diversos finais de semana de trabalho. Obrigada Ruth pelos saborosos almoços com sucos de abacaxi com hortelã e bolo de cenoura com cobertura de chocolate. Obrigada Profa por cada cafezinho que gentilmente me serviu!

Todas essas oportunidades, observações e acolhimento contribuíram para o desenvolvimento deste trabalho, construíram esta candidata à pesquisadora e, sem dúvida, influenciaram profundamente a pessoa por trás da aluna!

Obrigada, Profa, pela generosidade e privilégio do convívio tão leve e enriquecedor. É uma honra.

It is better a well-made head to a well-filled head
– Edgar Morin apud Michel de Montaigne

Resumo

CORDEIRO, Kelli de Faria. **aDApTA: Adaptive Approach for Information Integration to Support Decision Making in Complex Environments**. 2015. 148. Tese (Doutorado em Informática) – Instituto de Matemática, Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

Em ambientes complexos, o processo de tomada de decisão enfrenta vários desafios devido à dinâmica de eventos inesperados e imprevistos. O fluxo de informações é dinâmico e os dados são heterogêneos; em geral, não podem ser previstos o que impede seu carregamento prévio. Para prover visões confiáveis, integradas e atualizadas da situação em andamento, os recursos de informação são constantemente revistos. Nesse contexto, é necessário lidar com múltiplas fontes e domínios, além de incertezas sobre a interoperabilidade de dados entre esquemas heterogêneos. A rígida estrutura de bases de dados convencionais dificulta a adaptação dos sistemas de informação às demandas que não podem ser antecipadas em tempo de projeto. Como alternativa, os modelos de dados em grafo podem lidar com a heterogeneidade estrutural, complementados por representações semânticas como Linked Open Data (LOD) na Web. A ampla e crescente disponibilidade dessas fontes de dados agora representa um recurso importante para atender às demandas de informação que emergem no decorrer das mudanças no ambiente. Neste contexto, a adaptação pode ser considerada como uma forma de resolução de conflitos entre os elementos de um ambiente complexo. No âmbito de um sistema de informação, um elemento é um recurso de informação com características estruturais e nível de expressividade semântica que podem entrar em conflito com as características de outros recursos de informação, assim comprometendo a integração. Além da resolução desses conflitos, a abordagem e ferramentas utilizadas para a integração utilizadas para integrar as informações devem ser adequadas às características dos recursos de informação. Apesar dos esforços empregados em pesquisas para aprimorar a integração de informações, mesmo utilizando estruturas semânticas, escolher a abordagem de integração mais apropriada para o nível semântico até então desconhecido, ainda é uma questão em aberto. Além disso, a proveniência de todo o processo deve ser coletada para apoiar as avaliações de qualidade de informação. Com base neste cenário, esta tese propõe a aDApTA, uma abordagem

adaptativa para integração de informações com uma arquitetura associada. A aDApTA é apoiada pelos princípios de LOD usando um ETL workflow (Extract, Transform and Load) para as transformações, coletas e interligações tanto dos dados do domínio quanto dos dados de proveniência. A viabilidade da proposta é avaliada através de um protótipo usando fontes de dados de um cenário real sobre logística humanitária. Neste cenário, a logística de materiais de primeiros socorros é apoiada por uma visão dinâmica com múltiplas perspectivas. Os resultados sugerem que, embora os dados de entrada tenham heterogeneidade semântica e estrutural, uma visão integrada e com informações de qualidade podem ser construídas para apoiar a tomada de decisão.

Palavras-chave: Integração de Informação, Sistemas de Informação Complexos, Adaptação, Apoio à Tomada de Decisão, Dados Ligados, Web Semântica, Gestão de Emergências.

Abstract

CORDEIRO, Kelli de Faria. **aDApTA: Adaptive Approach for Information Integration to Support Decision Making in Complex Environments**. 2015. 148. Doctoral Thesis (Doutorado em Informática) – Instituto de Matemática, Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

In complex environments, the decision making process faces several challenges due to the dynamic of unpredicted events. There is a high dynamic flow of information where most of the relevant data are heterogeneous, cannot be predicted and loaded in advance, requiring constant revision of information resources to provide reliable, integrated and updated views of the situation. It is necessary to deal with multiple sources, domains, and uncertainties of data interoperability between heterogeneous schemas. The rigid structure of conventional databases makes it difficult to adapt the system to demands that could not be anticipated at design time. As an alternative, graph data models can handle structural heterogeneity complemented by semantic representations like Linked Open Data (LOD) on the Web. The wide and growing availability of LOD sources is now an important resource to meet the demands for information from decision makers throughout situational changes. In this context, we argue that adaptation can be seen as a way of solving conflicts between the elements of a complex environment. In the scope of an information system, a data resource is an element with varied structural characteristics and semantic expressivity, generating conflicts during the integration process. These conflicts must be solved to enable data integration. Once conflicts are solved, the integration approach must be suitable to the information resource characteristics. Even though many efforts have been made in data integration research, choosing the appropriate integration approach for an unknown semantic level of information is still an open issue. Moreover, the provenance of the whole process must be collected to support information quality assessments. Based on this scenario, this thesis proposes the aDApTA, an adaptive approach for information integration and its associated architecture. It is supported by LOD principles using an ETL (Extract, Transform and Load) workflow with provenance collection. The feasibility of the proposal is evaluated through a prototype using data sources of a real scenario about the humanitarian logistics. In this scenario, the logistics of relief goods are managed through a dynamic and

multi-perspective view. The results suggest that, even though the incoming data have semantic and structural heterogeneity, a reliable integrated view can be built to support decision making.

Keywords: Information Integration, Complex Information System, Adaptation, Decision Making, Linked Data, Semantic Web, Emergency Management

List of Figures

List of Tables

List of Acronyms

API	Application Programming Interfaces
ETL	Extract, Transform and Load
ETL4LOD	Extract, Transform and Load for Linked Open Data
ETL4LOD-Graph	Extract, Transform and Load for Linked Open Data Graph
LDIF	Linked Data Integration Framework
LOD	Linked Open Data
PDI	Pentaho Data Integration
RDF	Resource Description Framework
SPARQL	Protocol and RDF Query Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator

Contents

Introduction

1.1 Motivation

In complex environments, the decision making process faces several challenges because of the dynamic of unpredicted events. There is a high dynamic flow of information where most of the relevant data are heterogeneous, cannot be predicted, and cannot be loaded in advance. In addition, the reuse of third party data can be compromised without proper data quality and trustworthiness assessments (SNOWDEN; BOONE, 2007; BHAROSA; JANSSEN, 2010; BARR et al., 2011).

Among several definitions through different disciplines, in this thesis, we consider that a complex environment is composed of heterogeneous elements interacting with each other in an unpredictable way. An element can be an information resource to be integrated with another one. While supporting decision making, one more issue arises: the trustworthiness. Furthermore, we assume that dynamic is one aspect of a complex environment, and other aspects also include imprecision, antagonism, and uncertainty (SIMON, 1962; AXELROD; COHEN, 2000; BENINI et al., 2003; BENBYA; MCKELVEY, 2006; MORIN, 2007; BARRAT et al., 2008).

Over time, data integration approaches have been evolving to handle the dynamics of information generation and emergent demands (BLEIHOLDER; NAUMANN, 2008; BELLATRECHE et al., 2012; DE VETTOR, 2014). Several works have started to recognize the research heterogeneity issues, which causes conflicts and may prevent integration (LENZERINI, 2002; RAM; PARK, 2004; MICHELFEIT et al., 2014). In parallel, many works have proposed conflict resolution approaches. One of the most widely used is mediation where a mediation schema is mapped to the sources schema in order to enable integration (HALEVY, 2000; LÓSCIO, 2003; SALGADO et al., 2011; MICHELFEIT, 2013). In addition, ontologies have been widely used to improve the results of the integration process (KONDYLAKIS et al., 2009; JAIN et al., 2011; GAO, 2012).

More recently, to handle the information heterogeneity, especially when schemas are not known in advance, some works have adopted adaptive procedures to information

integration, considering this dynamic characteristic (KHAZANKIN; DUSTDAR, 2010; HEDELER et al., 2012; MORI; CLEVE, 2013; YONGTAO et al., 2013; SELLAMI et al., 2014).

Based on the literature review, earlier approaches for information integration mainly used traditional relational databases where schemas are defined a priori and data are inserted and maintained accordingly. Only data that complies with the structure and associated rules can be stored. In complex environments, the unexpected structure of the data that comes from different sources and organizations will most often not meet previously defined rules and structures. As an alternative, graph data models, complemented by semantic representations, such as Linked Open Data (LOD) on the Web, can handle heterogeneity, which we can find in many recent works (FERRARA et al., 2011; SCHARFFE; EUZENAT, 2011; WÖLGER et al., 2011; ARAUJO, 2014).

In this scenario, the volume of data published on the Web is growing constantly, especially motivated by government data and social movements (SCHMACHTENBERG et al., 2014). As a consequence, new kinds of information heterogeneity emerge. Data are published on the Web with different types of descriptors, some with a lack of semantic concerns and others with high semantic concerns. Besides the wide range of heterogeneity levels identified by research in literature and the wide problems covered by solution approaches (LENZERINI, 2002; RAM; PARK, 2004; SCHULTZ et al., 2011; MICHELFEIT et al., 2014), there is still a lack of adaptive features to handle heterogeneous semantic expressivity levels.

1.2 Problem

From a general perspective, the problem addressed in this thesis is how to support the integration of information for decision making in complex environments where most relevant data are heterogeneous and cannot be known in advance. More specifically, how to provide a reliable and integrated view of the situation having to accommodate information from unpredicted sources? In addition, how to identify and help to solve heterogeneity conflicts during the integration process? Accordingly, there is still a lack of adequate methods and tools to handle the dynamics of information flow in complex environments, even though sound results have been achieved in several works about information integration.

1.3 Hypothesis

The hypothesis to solve the described problem is that an adaptive approach, which considers characteristics of the unpredicted sources, may improve the integration with a flexible process and with the generation of more expressive relations between the resources. In this thesis, an adaptation is considered as a way of resolving conflicts between the elements of an environment by changing the components' behavior in order to restore the system flow. In complex systems, an element can change its behavior in an unpredicted way, causing the interruption of the interaction flow. In the scope of an information system, an element is an information resource with characteristics that can conflict with the characteristics of an element of another information resource, preventing proper integration. Thus, more specifically, the hypothesis is that the mitigation of the heterogeneity conflicts may enable information integration and improve its results.

1.4 Solution approach

Based on the presented open issues and hypothesis, the solution approach is the adaptation of the system by changing the integration method according to the information characteristics. Based on Schonenberg et al. (2008) taxonomy of flexibility, this approach can be classified as *flexibility by change,* which means the ability to modify a process allowing the adaptation to changes that are identified in the operating environment.

The heterogeneity levels of the information resource are identified and the more appropriate integration approach is chosen and applied. The approach also supports different semantic treatment according to semantic expressivity of data sources. An adaptive support can facilitate and stimulate the semantic enrichment of data with elementary descriptions, in addition to the application of more sophisticated conceptual models with explicitly ontological commitment (CORDEIRO et al., 2011a).

Thus, the approach, called aDApTA, adapts the integration process based on the semantic expressivity level of data sources with corresponding provenance data collection. Supported by an Extract, Transform and Load (ETL) workflow, aDApTA uses Resource Description Framework (RDF) Graph representation, based on LOD principles, as a strategy

for data structure and description adaptation. Furthermore, the provenance of the integration workflow supports quality assessment.

1.5 Objective

This research aims to provide an approach for adaptive information integration to support decision making in complex environments. To achieve that, the specific objectives are: (i) to propose a set of activities to identify the heterogeneity levels of unpredicted data sources and to allow the adaptation of the integration approach accordingly; (ii) to provide a supporting architecture with core components enabling the approach implementation in any domain; (iii) to enable trustworthy assessments of the integrated information; and (iv) to present a prototype and case study to show the approach application.

1.6 Methodology

To achieve the described objectives, the research method follows these phases: (i) identification of the main characteristics of information used to support decision making in complex environment focusing on its dynamic flow; (ii) exploration of the approaches published in literature, used to integrate heterogeneous and unpredicted data sources; (iii) identification of the heterogeneity types and levels distinguished by the reviewed research works, highlighting the missing level, semantic expressivity; (iv) study the adaptive features used in the approaches which handle information conflicts in dynamic environments; (v) modeling a conceptual framework to support the description of information semantic expressivity level and the corresponding integration approaches; (vi) formulation of an approach to identify and to solve structure and semantic expressivity conflicts enabling the selection and switching of the interlinking method; (vii) design of an architecture and a development framework to support the approach application in different domains; (viii) use data provenance collection to support information quality and trustworthy assessments; and (ix) implementation of a prototype.

To enable the approach evaluation, aDApTA was applied to data sources of a real scenario about the humanitarian logistic operations of the Typhoon Haiyan disaster in the Philippines. Humanitarian logistics is a typically complex environment where the analysis of an integrated dataset can produce non-trivial results as external data acquisition is

unpredictable and additional uncertainty must be considered (BENINI et al., 2003; BHAROSA; JANSSEN, 2010). The feasibility of the proposal was evaluated through the development and application of a prototype for the Philippines case. The initial research results suggest that even though the incoming data have semantic and structural heterogeneity, an integrated view can be built by using appropriate interlinking approaches. Qualitative and quantitative measures were developed to show that when better links are created, the data are more expressive. Moreover, the results showed how the collection and publication of fine-grained provenance data of the ETL workflow can support a trustworthy data assessment.

1.7 Structure

This thesis is organized as follows:

Chapter 2 identifies issues of information management in relation to complex environments, characterizing the heterogeneity levels faced and the corresponding conflict resolution approaches, highlighting some of the issues that are still open in current research works.

Charter 3 provides general concepts of graph data models and the semantic representations used in the solution approach. In addition, some tools available in the literature for information integration, which uses graph data models to improve the results, are described.

Chapter 4 presents the core of this thesis, the proposed approach called aDApTA that handle the information integration issues through a set of activities organized in a process flow. In addition, to enable the approach implementation in any domain a supporting architecture and a development framework is described.

Chapter 5 explores an application case used to highlight the contributions of the thesis and the feasibility of aDApTA using data from a real case scenario and discussing the results.

Chapter 6 finally concludes the thesis describing the contributions, limitations, and future works for this research.

2 Information Management in Complex Environments

The complexity of external environments presents a major challenge for the design of information systems that support decision making. Aspects of complex environments are discussed in some theories that provide conceptual frameworks as ways of thinking and seeing the world. Some philosophical and social theories also address these complexity aspects in order to understand them, acknowledging the difficulties for eliminating or controlling them. Those theories seek ways to deal with complexity features, such as unpredictability, dynamics, imprecision, and heterogeneity. Among the approaches to handle these features there are adaptation, integration, and feedback (SIMON, 1962; AXELROD; COHEN, 2000; BENBYA; MCKELVEY, 2006; MORIN, 2007; SNOWDEN; BOONE, 2007; BARRAT et al., 2008; MITCHELL, 2009).

According to Axelrod and Cohen (2000), “A complex system is one in which the actions of agents are tied very closely to the actions of other agents in the system.” Based on this and other theories about complex systems, Benbya and McKelvey (2006) suggest that “information systems should not be developed as static entities.” These positions are aligned with the fact that, in recent years, the increasing dynamics of the external environment impose high rates of change in information systems, especially those that support activities at the strategic level. Activities at this level have greater interaction with the external environment and process a higher volume of data, compared to activities at the operational level.

In a typical complex scenario, a system stores information representing different types of knowledge, as illustrated in Figure 1. Previous knowledge comprises known information, and current knowledge refers to the situation awareness described by information with heterogeneous and unpredicted structure. Current and previous knowledge are integrated to compose what had been called combined knowledge to support decision making (DINIZ et al., 2005).

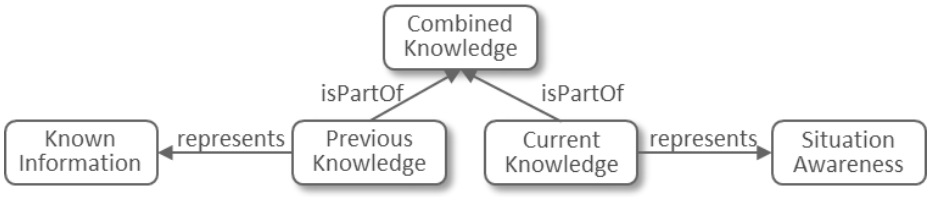


Figure 1: Conceptual map of knowledge support for decision making. Adapted from (DINIZ et al., 2005)

To compose the combined knowledge, the integration task handles a high dynamic flow of information, which has some particular characteristics and problems that need to be understood, analyzed, and addressed.

2.1 Dynamic of Information Flow in Complex Environments

In this section, the problem addressed in this thesis is characterized through our view of the dynamic of an information flow in complex environments. This view was built upon several works available in literature; some of them are: Diniz et al. (2005), Benbya and McKelvey (2006), Halevy et al. (2006), Nicolle and Cruz (2009), Bharosa and Janssen (2010), Barr et al. (2010, 2011), and Santos et al. (2011). One of the main issues addressed in these works is the influence of complexity aspects in building the combined knowledge to support decision making.

A starting point for understanding the dynamics of an information flow in complex environments is characterizing its unit. A piece of information or an *information resource* is composed by an *instance* (the data itself) and its corresponding description, the *schema* or *structure*. An information resource can be described depending on the format. For example, it can be the label of a column head of a tabular data, the tag label of an XML document, the table and column names of a relational database, the label of vertices and edges of a graph data and more. Figure 2 depicts the dynamic and the characterization of information types and flows; an information resource and its schema are represented by diagonal lines and a rectangle border, respectively.

An information resource can represent *previous knowledge,* which comprises known information and can be structured in order to be reused in future situations. The schema and instances representing *previous events* are loaded from agents handling the events on the field (Flow 1) or from worldwide available datasets about the events (Flow 2).

On the other hand, the structure of *current knowledge* can only be partially predicted increasing the difficulties of information management. Initially, it is possible to predict that some types of information will be needed. In this case, they can be designed on the information base schema before the event occurs. Therefore, one might define the schema a priori, but only during the actual situation where the data will be available (Flow 3). However, a set of information (instances and schemas) representing current knowledge can come in unpredicted ways from unexpected reports about the ongoing situation with unpredicted structures, which is, therefore, characterized as *unpredicted information*. Also, complementary information to meet emergent demands can be gathered from unpredicted sources available on the Data Cloud.

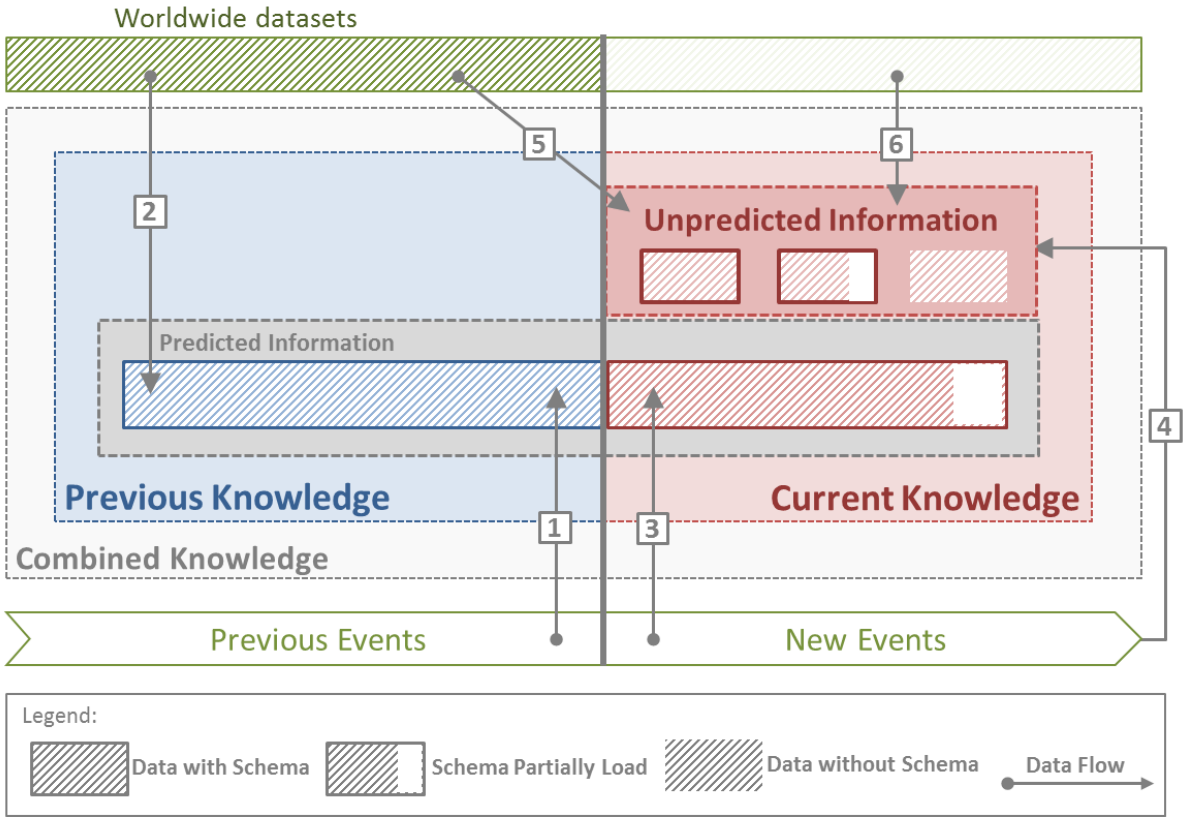


Figure 2: Dynamics of information flow in complex environments

Unpredicted information can arise with or without a schema. Moreover, instances can arise with a schema that can accommodate more instances during event handling. For example, in an emergency management situation, the available storehouses can be informed with some properties, such as location and capacity, in a spreadsheet partially

fulfilled. Later, new attributes can be added, such as the manager. During the event, more storehouses can be built, requiring the input of more instances. There are also other possibilities, such as information partially structured or with insufficient descriptions, i.e., with different levels of expressivity. For example, the locations in a disaster manager are grouped in sectors, called *arbitrary regions*. The information about these regions can be informed using standard vocabularies with many properties or simply described as a cluster without any others properties. This particular characteristic represents a major challenge while composing the combined knowledge in a dynamic environment because the current knowledge can be informed by field agents with different descriptors and levels of semantic expressivity. This issue is explored in this work.

The sources of unpredicted information can be of several types, such as reports about the ongoing situation (Flow 4), or datasets related to the event, available worldwide with open access, such as the Web of Data (Flow 5). In this last particular case, data can exist on the Web and be retrieved as necessary during event handling to meet unpredicted demands of information. Accordingly, the information created before the event becomes part of the current knowledge. Additionally, some datasets on the Web can be created during event handling (Flow 6).

In short, unpredicted information about the situation represents contextual information with or without an associated description. They are mostly from a domain that cannot be anticipated. Furthermore, this unpredicted information is mainly characterized by different semantic expressivity levels. A low level of expressivity is expected due to the dynamics imposed by the complex environment. On the other hand, an unpredicted demand for complementary information usually corresponds to data with a schema, retrieved through querying data sources, from a known domain. Usually, different semantic expressivity levels might be found, especially on the Web of Data, but the higher levels are expected considering that the information was generated in a scenario where the time pressure was not so critical.

Several research works handle these problems with different strategies and also focus on different parts. In order to understand them and highlight the gaps, some complexity aspects of information management were factored, initially, by heterogeneity,

unpredictability, and imprecision issues, and some solution works are presented accordingly in next section.

2.2 Information Integration from Heterogeneous and Unpredicted Sources

Over time, to compose the combined knowledge view, the decision making process requires information integrated from an increasing number of heterogeneous data sources that can emerge dynamically. In the past 30 years, there were several research proposals addressing information **integration,** which includes a variety of terms, such as mediation, mapping, matching, alignment, interlinking, fusion and more. In some of these works, the **heterogeneity** of information is distinguished through syntax, structural, and semantic levels. The **unpredictability** of data sources, one of the main issues in a complex environment, has being addressed by adding adaptive features to the information systems. Additionally, the use of third party sources of data brings one more issue, that of **trustworthiness**. Recently, as the variety of data sources has become distributed on the Web, the problem of data quality has received special attention. The problem has been addressed through the collection of provenance data. Some of these works are described in the following sections.

2.2.1 Information Heterogeneity Levels

Ram and Park (2004) classify the information heterogeneity conflicts in syntax, schema, and semantic levels. Syntactic heterogeneity is caused by the use of different models or languages, such as relational and XML. Schematic heterogeneity results from structural differences where an entity can be related to others through an attribute or a relationship. Semantic heterogeneity is caused by different meanings or interpretations of terms in various scenarios. Similarly, De Vettor et al. (2014) also classified the information heterogeneity issues along three levels: syntactic, i.e., related to data formats and syntax; structural, due to differences in data organization; and semantic, when different knowledge representations are used.

Bellatreche et al. (2012) suggested that any integration system should take into consideration the resolution of syntactic, schematic, and semantic conflicts. Along those lines, in order to achieve complete integration of heterogeneous data, Gao (2012) stated

that different integration tasks are required. On the syntactic level, the integration task solves the different data type problems, e.g., short integer versus long. In the structural level, the resolution is performed by re-formatting the data structures to a homogeneous data structure, such as XML. The semantic level is about the meaning of terms in a special context or application. It refers to different semantics due to different conceptualizations of similar concepts.

Exploring different types of conflicts, Michelfeit et al. (2014) detailed the classification in schema, identity, and data conflicts. Schema conflicts are caused by different source data schemata—different attribute names, data representations (e.g., one or two attributes for name and surname), or semantics (e.g., units). Identity conflicts are a result of different identifiers used for the same real world objects. Data conflicts occur when different conflicting values exist for an attribute of one object. Conflict can be resolved on the entity or attribute level by a resolution function. Resolution functions can be classified as deciding functions, which can only choose values from the input such as the maximum value, or mediating functions, which may produce new values such as the average or sum. This classification is in accordance with Bleiholder and Naumann (2008) who also classified the conflicts in three levels. First, there are schematic conflicts, such as different attribute names or differently structured data sources. Second, there are identity conflicts, as the way of identifying a real-world object may be different in the data sources. Data conflicts are caused by multiple representations of same real-world objects.

As described, the heterogeneity levels are common sense in literature of information integration, even though the categories may differ in terms and concepts. This problem has been addressed using conflict resolution approaches (BROWN; DUREN, 1986). Bleiholder and Naumann (2008) classified the strategies to handle conflicts, as illustrated in Figure 3.

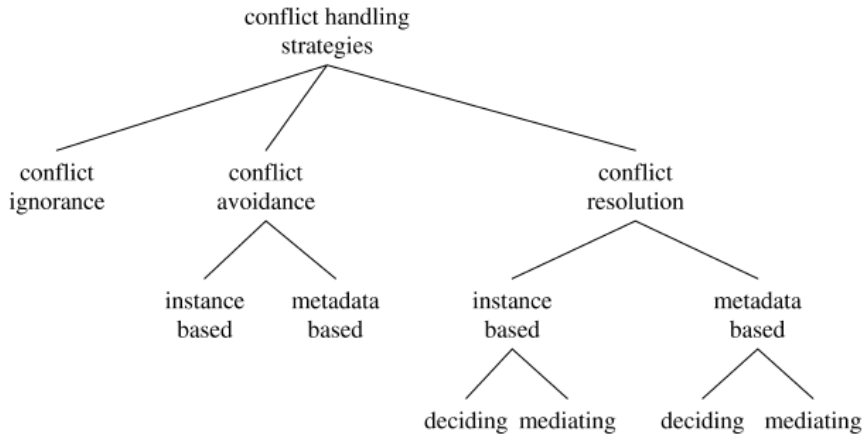


Figure 3: Conflict handling strategies (BLEIHOLDER; NAUMANN, 2008).

The conflict resolution is performed in different ways in information integration systems. A broad range of strategies has being developed to solve a conflict throughout the levels. Focusing on the semantic heterogeneity level, some works use semantic measures (SM) to solve the conflict (VOLZ et al., 2009; ARAUJO, 2014). According to Harispe et al. (2013a, 2013b), SMs can be used to compare various types of elements: units of language: words, sentences, paragraphs, documents; concepts/classes, groups of concepts; and, instances semantically characterized. This last one can be the RDF description of the corresponding instance, a set of conceptual annotations associated to it, a set of tags, or even a subgraph of an ontology. The authors introduced and classified the large diversity of approaches of semantic measures, as illustrated in Figure 4.

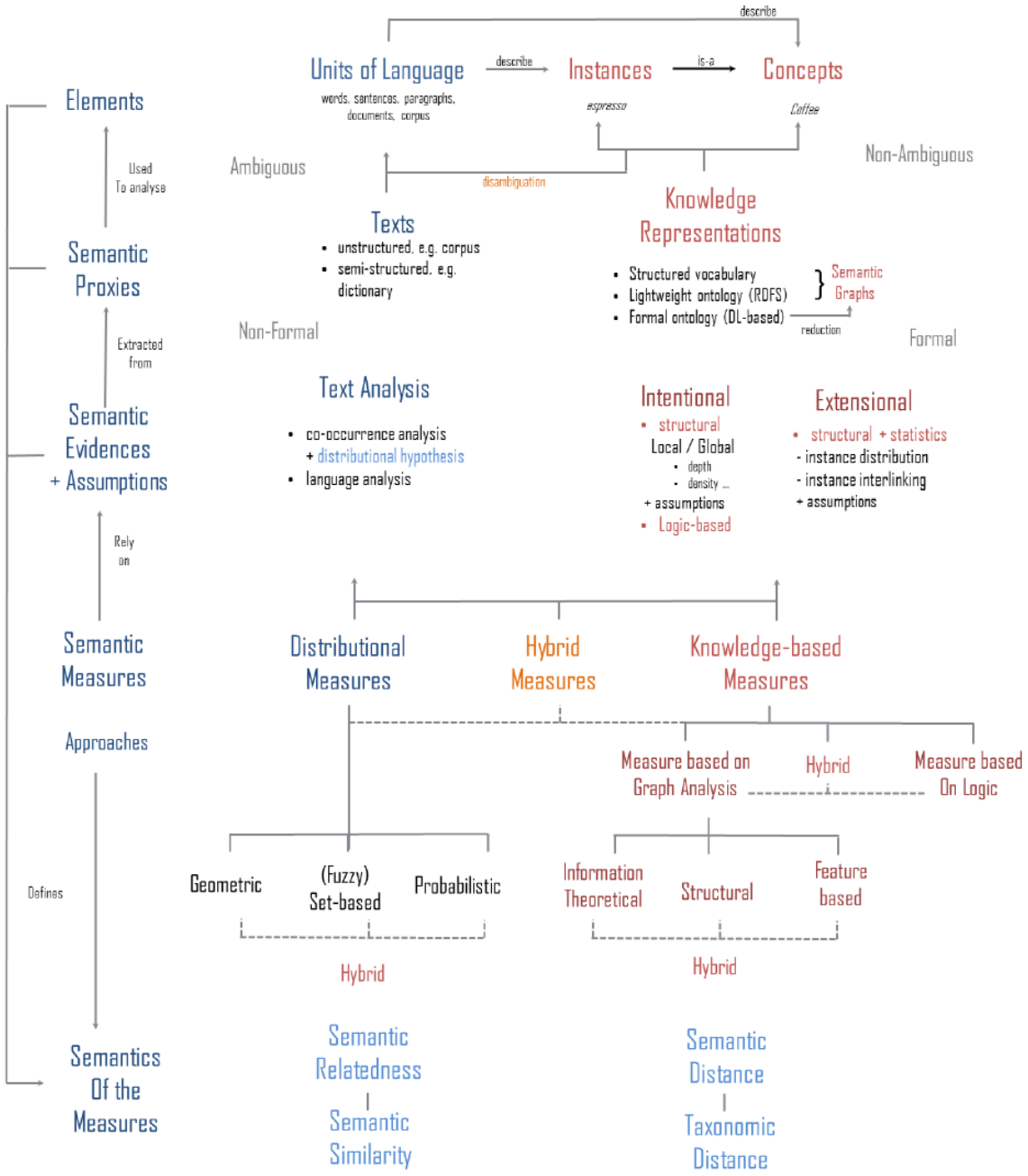


Figure 4: Types of semantic measures (HARISPE et al., 2013a)

2.2.2 Integration of Information

Data integration can be defined as the problem of combining data from different sources or can provide users with a uniform view over a set of heterogeneous and distributed data sources (LENZERINI et al., 2002; SALGADO et al., 2011). The whole process of data integration is a complex task with many subtasks, and has been studied mostly in relation to relational databases (HALEVY, 2000; MICHELFEIT, 2013). More recently, several works used other information representations, like ontologies and Semantic Web

technologies to improve the results of the integration process (LENZERINI, 2002; BRANDAO et al., 2010; SALGADO et al., 2011; SILVA et al., 2011; GAO, 2012; ARAUJO, 2014; MICHELFEIT et al., 2014). The processes may largely vary on methods, system, scope, and terms, which can cause confusion in comprehension and gap assessments. However, most of them focus on the semantic level conflict resolution, as it can be considered the most difficult integration task.

According to Oxley and Thorsen (2004), more generally, integration is the process of connecting systems into a larger system, and these systems may have **fusion** tasks implemented. These authors consider fusion as the process of mapping several objects onto a single object in an optimal fashion. Based on the same argumentation, Bleiholder and Naumann (2008) describe data fusion as the task of fusing multiple records representing the same real-world object into a single, consistent, and clean representation. In Mendes et al.'s (2012) proposition, the fusion can be performed by solving conflicts between objects values defining rules, such as by maintaining the max or min value, or calculating and maintaining the medium. Before the fusion, a mapping is implemented in order to eliminate the duplicates. Oxley and Thorsen (2004) differentiate integration from fusion: "Integration can be defined without having to specificy the systems' elements. Fusion is the process of mapping several objects to a single object in an optimal fashion. Integration is the process of connecting systems into a larger system. These systems may have fusion in them." Over time, these two concepts have been employed distinctly; fusion can be considered one of techniques used to integrate information.

While fusion focuses on object values, **mediation** approaches focus on mapping databases' schemas. The mapping can be created manually, automatically or semi-automatically. The approach may be implemented using as a global schema expressed as queries over sources, called Global-As-View (GAV). In an opposite way, each object in a given source can be defined as a view over the mediated schema, called Local-As-View (LAV). Moreover, combining them is also possible, called Global-Local-As-View (HALEVY, 2000; LÓSCIO, 2003; SALGADO et al., 2011; MICHELFEIT, 2013).

In the ontologies area, Gao (2012) reviewed many integration concepts used in literature. **Mapping** can be considered as a set of formulas that provides the semantic

relationships between the concepts in the ontologies. **Alignment** represents the mapping of concepts and relations between multiple ontologies based on preservation of the partial ordering and synonyms, as well as the possible introduction of new concepts that will function as sub- or super-types. **Matching** can be considered as the correspondence between individual concepts of the two ontologies or the identification of maximal one-to-one correspondences between elements (e.g., concepts). It can be performed by finding the correspondence of terms based on a set of rules, to create the intersection. Furthermore, matching enables the analysis of similarities and differences between the concepts to predict their semantic compatibility. According to Gao (2012), some works have used mapping and matching almost interchangeably, though the authors also noted that matching is used for the automated part while mapping is for the overall process that usually involves human intervention. Based on these concepts, mediators and mapping algorithms are used to match similar concepts. Some mathematical and logical tools are used to infer and fuse information. The result is a set of aligned information based on a well-known ontology (ARAUJO et al. 2010; BELLENGER, 2013).

In the Web of Data area, the information resources are connected by **interlinking** individual instances and schemas, in a process usually referred to as instance matching and schema matching, respectively. Araujo (2014) defines instance matching as “the problem of finding two or more distinct instance representations that refer to the same real world entity.” Jain et al. (2010) present a system for finding schema-level links between LOD datasets using ontology alignment methods. Assaf et al. (2012) improved schema matching techniques using LOD. In this scenario, the result of the matching process can be a set of links pointing to the information resources with an identified relationship (ISELE et al., 2010). Moreover, the interlinking can focuses on data integration applying fusion techniques through conflict resolution and quality assessment (MENDES et al. 2012; MICHELFEIT, 2013; ARAUJO, 2014).

Regardless the broad concepts used to represent the integration methods, what can be noticed throughout the evolution of the research works is that the methods are being reused and improved within other methods as new issues emerge. Lately, due to the

growing sources of data, the dynamics of information flow are imposing great modifications on application of integration methods, as explored in following section.

2.2.3 Handling Unpredicted Data Sources

Many research works have addressed heterogeneity, a common complexity aspect, distinguishing the conflict levels and proposing corresponding solutions. However, one more aspect of a complex environment represents a great problem—the unpredictability. To handle unpredicted issues, following the suggestion of Benbya and McKelvey (2006), an information system "should be allowed to grow and adapt to emergent user requirements." Axelrod and Cohen (2000) state, "When the agents in a system are actively trying to improve themselves ('adapt'), then the system is a Complex Adaptive System." Thus, to generate combined knowledge in a complex environment, the information management system must adapt to emergent requirements. According to Bharosa and Janssenm (2009), information management adaptability can be defined as the ability to rapidly change existing resources or create new ones in order to align the internal information demand with external information supply and events. The same authors also define adaptivity as the ability to maintain the homeostasis, i.e., the dynamic equilibrium in an open system, critical to the system's function.

Following these concepts, to handle semantic conflicts in dynamic environments, many works have proposed approaches based on **schema evolution** and mapping maintenance. Others have proposed systems with **architectural components**, such as agents, middleware, mediators, and wrappers to adapt the integration process to the ongoing data sources characteristics. Furthermore, some of them act on a **query** processing layer.

Bouguettaya et al. (2003) used **agents** to allow adaptive intercommunity relationships in WebFINDIT, an adaptive Web-based database. WebFINDIT provides a monitoring mechanism to alter dynamically relationships between different database communities. This is achieved by using distributed agents that work as background processes. Their role is to continually gather and evaluate information about the intercommunity relationships to recommend changes. They use the notion that agents are software components characterized mainly by their autonomy and adaptiveness (NWANA;

NDUMU, 1997). On the same basis, Thollot (2012) focused on dynamic situation management, which is how to continuously monitor agents' interactions and help maintain situation graphs as events occur in their environments. They thus define two types of active components to enable dynamic maintenance of situation graphs: activation rules and operators.

Using **architectural components**, Khazankin and Dustdar (2010) proposed an adaptive integration approach that allows the switching of the integration technique to the most appropriate one according to specified criteria. The authors focused on non-functional concerns, such as data volume, network latency, and throughput. The data from the Web is provided to the end application via the "proxy" source that acts as a wrapper using one of the underlying integration techniques. The technique swapping is controlled by the adaptation module that makes decisions based on the defined criteria. The control module checks periodically if the currently used technique is still the best. Langegger et al.'s (2008) integration proposition used a middleware based on mediators and wrappers over LOD datasets. Data heterogeneity is addressed by RDF-wrappers as a D2R-Server placed on top of local information systems. De Vettor et al. (2012), Mrissa et al. (2013), Sellami et al. (2014), and De Vettor et al. (2014) proposed the Data Mediation as a Service (DMaaS) approach, which is an automated solution for resolving data heterogeneity problems between semantically described data, using a decentralized (peer to peer) repository of mediation services. The DMaaS approach classifies data heterogeneity issues according to the syntactic, structural, and semantic levels, and provides adapted mediation along these levels. The approach sets up the automatic conflict detection mechanism, which analyzes input metadata, and intercepts data responses to perform reconciliation. Using the emergency management domain, HUODINI architecture (FAHLAND et al., 2007) integrates several sources of information available on the Web through the use of wrappers that transform data into RDF. In this architecture, the space and time properties of information are processed to support the representation of emergency dynamics situations. Yongtao et al. (2013) and Yongtao (2014) proposed *typification*—an approach to infer the type of semantics of structured data. The authors argue that instance matching solutions may not perform well when the type of information is either missing or too general to be useful.

Schema versioning is one of a number of related areas dealing with the same general problem—that of using multiple heterogeneous schemata for various database related tasks (RODDICK, 1995). Jun et al. (2002) supported an adaptive evolution of a schema in the form of expansion with new classes and/or compaction by removing inefficient ones. Jiacai et al. (2012) proposed an adaptive database schema design method for multi-tenant applications. Mori and Cleve (2013) presented a feature-based approach to adapt database schemas according to the changing context conditions. The method allows the derivation of a consistent and sufficient sub-schema starting from the current context. In this thesis, context means "what surrounds the focus of attention" and can be defined as "a complex description of shared knowledge about physical, social, historical, or other circumstances within which an action or event occurs", according to Borges et al. (2005, 2007). This concept is based on Brézillon and Pomerol (1999) who distinguish the part of the context being relevant for the current performer's focus of attention from the irrelevant part.

Salgado et al. (2011) stated that in a dynamic environment, mapping maintenance is an important task concerning data integration systems. Mappings between the mediated schema and the source schemas must be flexible enough in order to accommodate schema evolution. Each change at the source schema level may lead to the reconsideration and possibly the change of a set of existing mappings (mediation queries). The problem is addressed by Lóscio (2003) who proposed an approach for propagating a change event occurring at the source level into the mediation level, in such a way that the mediation level may evolve incrementally, and modifications can be handled easier, thus increasing the system flexibility and scalability. The use of ontologies with schema evolution for data integration was explored by Kondylakis et al. (2009).

The principles of Data Spaces proposed by Franklin and Halevy (2005) originated the Dataspace Platforms Support (DSSP), which adapts its integration mechanisms to heterogeneous data sources. In this platform, the query returns the best result as possible without guarantees of quality and accuracy (HALEVY et al., 2006). Moreover, the identification of semantic relationships among data makes use of human attention, in an approach that became known as "pay as you go." Based on this approach, the architecture PAYGO was proposed by Madhavan et al. (2007), focusing on mechanisms for on demand

integration using the users' feedback about the relationships between information. This feedback is used to rank future queries, thereby increasing the accuracy of response. Based on the same approach, Salles et al. (2007) proposed the iTrails for information integration in data spaces. Through the compilation and refinement of data space approaches and principles, the DSToolkit supports the entire lifecycle of data spaces: initialization, use, and enhancement. The architecture is based on the treatment of schemas represented in the heterogeneous models, on the treatment of the users' feedback about the query results and new data sources. The system is based on model management which uses operators for schema integration and evolving schemas (HEDELER et al., 2012).

To provide an integrated view, some works deal with the dynamics and distribution of data sources in the **query layer**. Le-Phuoc et al. (2011) proposed the Continuous Query Evaluation over Linked Streams (CQELS), which provides a flexible query execution framework with the query processor dynamically adapting to changes in the input data. During query execution, it continuously reorders operators according to some heuristics to achieve improved query execution in terms of delay and complexity. Lynden et al. (2010) describe an approach based on distributed query processing. Data from multiple repositories are used to construct partitioned tables that are integrated using an adaptive query processing technique. The approach supports a join reordering, which limits any reliance on statistics and metadata about SPARQL endpoints, as such information is often inaccurate or unavailable, but is required by existing systems supporting federated SPARQL queries. In addition, Acosta et al. (2011) proposed the ANAPSID, an adaptive query engine for SPARQL endpoints that adapts query execution schedulers to data availability and run-time conditions. ANAPSID provides physical SPARQL operators that detect when a source becomes blocked or data traffic is burst, and opportunistically, the operators produce results as quickly as data arrives from the sources.

2.2.4 Information Trustworthiness

Third party data sources are a vital element to meet unpredicted demands of information to support decision making in a complex environment. The data from these sources can be retrieved or can arrive with heterogeneity on different levels and be integrated through one of the approaches presented in previous section. However, the

information is useless if it is not reliable and accurate. Thus, to enable the use of third party data, provenance becomes of fundamental importance. Provenance comprises all information associated with the domain data, describing who, how, when, and why it was published. It enables the enrichment of the context surrounding the data supporting the assessment of data reliability and quality (CORDEIRO et al., 2011a; MARJIT et al., 2012). In several domains, data quality evaluation depends on provenance availability (BUNEMAN; DAVIDSON, 2010). For instance, official governmental data is the primary source for all levels of decision making in public administration as well as for citizen awareness and participation on government actions; business analysts rely heavily on data transparency to ensure data quality to support mission success; banking and financial systems become untrustworthy and fragile without policies for quality control and evaluation.

There is already a huge volume of work on provenance. It can be described in various terms depending on the domain where it is applied (VOLZ et al., 2009; SCHULTZ et al., 2011). It provides important documentation that is essential to preserve the original data product, to determine its quality and authorship, to reproduce, as well as to interpret and validate the associated results. One of the most important initiatives about provenance is the Open Provenance Model (OPM) (MOREAU et al., 2011.), which includes the basic building blocks (concepts) to represent provenance: artifact (an immutable state of an object), process (action taken on an artifact), and agent (entity that may facilitate, control or somehow influence a process). Other important models are Provenir (SAHOO et al., 2008) and PROV-DM (www.w3.org/TR/prov-dm/). The concepts defined by OPM are also present in these models (not necessarily named equally). However, they may go deeper, specializing some of the OPM general concepts. Provenir, for instance, distinguishes input data and parameters. PROV-DM provides a richer (including constraints) and more formal representation. It is now a W3C candidate recommendation. Researchers behind these initiatives are now working together on PROV-DM towards the standardization of a general provenance model.

Another important effort towards unification of the different views of provenance is the taxonomy presented by (DE LA CERDA; CAVALCANTI, 2012). It covers provenance issues from several different areas of computer science, such as scientific workflows, business processes, databases, distributed environments, and Semantic Web. The taxonomy is an

endeavor to classify the characteristics of provenance that primarily (a) captures architectural styles of provenance systems; (b) considers the moment that provenance data is captured by a given mechanism; and (c) considers the existing methods to store and access provenance data.

The provenance concepts have been applied in the information integration domain with different approaches. Mendonça et al. (2013) proposed an approach for generating and capturing provenance, which covers preparation and format transformation up to the publication of the integrated dataset. Michelfeit et al. (2014) focused on data fusion and conflict resolution proposing data fusion algorithms with provenance tracking and quality assessment of fused data.

2.3 Open Issues in Information Integration in Complex Environments

Heterogeneity and unpredictability are aspects presented in complex environments that cause a strong impact on information integration systems compromising the quality of the results. Moreover, due to the growing sources of data available on the Web, imprecision is another complex aspect with an increasing impact on information management, treated as reliability, trustworthiness or quality of data. To handle these complex problems, adaptation and integration are mainly used by the approaches in the literature. In order to analysze and enable the gap assessments, some research presented in previous sections was grouped by the complexity aspects and type of solution approach, as organized in Table 1. They were chosen considering their overlap with the characteristics of the problem addressed in this thesis. Notice that the solutions may have different focus and integration approaches. However, most of them address some information heterogeneity level conflicts, which emerge in an unpredicted way, with adaptive solutions.

Despite the consistent distinctions between heterogeneity levels and sound results achieved by approaches that handle conflicts on each level, there is still a missing type of heterogeneity, deeply correlated to the semantic level—the expressivity. While semantic heterogeneity concerns different conceptualizations to the same real world object, the semantic expressivity heterogeneity concerns the different description resources used to make explicit the object's meaning. The expressivity level refers to the measure of how a data model describes the reality or the expressivity needed for a better description of real

world phenomena in models (SALTOR et al., 1991). This type of heterogeneity can prevent semantic integration of information, especially when an information resource has insufficient descriptions. For example, Yongtao et al. (2013) extract three million entity descriptions from several datasets crawled on the Web, and found that almost four hundred thousand of them lack type information—i.e., the type of these entities is not known. The authors state that the type-specific integration solutions do not perform well when the type of information is missing.

Table 1: Complexity aspects and solutions strategies of research works

			Main Related Works						
			Khazankin and Dustdar (2010)	Halevy et al. (2006) and Hedeler et al. (2012)	Mori and Cleve (2013)	De Vettor et al. (2012; 2014), Mrissa et al. (2013), Sellami et al. (2014)	Yongtao et al. (2013), Yongtao and Tran (2013), Yongtao (2014)	Knap et al. (2012), Michelfeit (2013), Michelfeit et al. (2014)	This Thesis
			-	DSToolkit	-	DMaaS	TYPifier	ODClenStore	aDApTA
Complexity Aspects	Heterogeneity Levels	Syntax	-	-	-	X	-	-	-
		Structure	-	-	X	X	-	-	X
		Semantic	-	-	X	X	-	X	X
		Schema	-	X		-	X	-	-
		Data	-	-	-	-	-	X	-
		Identity	-	-	-	-	-	X	-
		Semantic Expressivity	-	-	-	-	-	-	X
	Trustworthiness		-	X	-	-	-	X	X
	Unpredictability		X	X	X	X	X	-	X
	Uncertainty		-	X	-	-	-	-	-
Solution Strategy	Adaptation		Evaluation and Switching Components	Pay-as-you-go with feedback	Contextual Schema Derivation	Evaluation and Mediation Services	Evaluation and Typification	-	Evaluation and Switching Component
		Focus	Non-functions requirements	Query Layer	Mobile Application	Query Layer	Semantic Requirements	Semantic Requirements	Semantic Requirements
	Integration		Federation, Consolidation, Propagation	Model Management	Mapping	Aggregation and Merge	Instance Matching	Fuse	Interlinking Approaches
	Provenance		-	-	-	-	-	X	X

Furthermore, the importance of distinguishing this type of heterogeneity is growing because of the increasing ways to describe information, especially that which is published on the Web. As shown in Table 1, there is still a lack of approaches to characterize the semantic expressivity levels of an information resource, to identify the conflict, and to solve it.

Besides the heterogeneity concerns, in a complex environment, the information characteristics may emerge in an unpredicted way. Some works have addressed this issue monitoring some criteria variables and changing the integration approach accordingly. Among others, this approach presents a simple and natural way to deal with the unexpected. The changing of the integration mechanism is performed by Khazankin and Dustdar (2010) using a switching component, and by Sellami et al. (2014) using a mediation service. However, the first work addresses non-function requirements, and the second acts on query layer and does not store the generated connections. Also, with an evaluation approach, Yongtao et al. (2013) try to improve the integration results, by discovering the missing type of information.

Although the broader issues covered by some of the related works, a solution has not been found to handle the presented conflicts together with the missing heterogeneity type under a dynamic environment. Hence, in this scenario, there are unsolved problems, such as how to integrate information with different semantic expressivity levels, different organization structures, or file formats without knowing these characteristics in advance. How to identify the heterogeneity levels and switch the integration approach accordingly? Furthermore, how to support data quality assessments within heterogeneous and unpredicted issues?

Due to the schema rigidity of relational databases, more flexible representations, such as those supported by graph databases, can be used to address the presented problems since the absence of structuring rules does not prevent data from being stored, associated with other data, and even described. Hence, graph data models are more suitable to accommodate data items, annotating them through new links. This type of data structure is based on the proposed information base adaptation approach and explored in the following chapter.

3 Semantic Representation in Graph Data Models

Graph data models can be used as a strategy for data structure and semantic adaptation. Formally, a graph or property graph model is a directed graph composed of vertices and directed edges described by properties representing concepts (see Figure 5a and Figure 5b). In graph databases, there is no structuring schema. Data can be stored without identification of its defining type. The identification can be done later by adding new edges and vertices, when appropriate (ROBINSON et al., 2013). In this case, an edge can represent the relationship between an entity and a literal value (see Figure 5b). Also, an edge can represent the relationship between two vertices representing two entities.

New relationships can increase the semantic expressivity by adding new attributes to the entity. Figure 5c illustrates a graph data example about a location. The entity *Central Visayas*, one of the regions affected by the Haiyan typhoon disaster in the Philippines, is described by the entity *Place,* has name "Region VII" and has a population of 6,800,180.

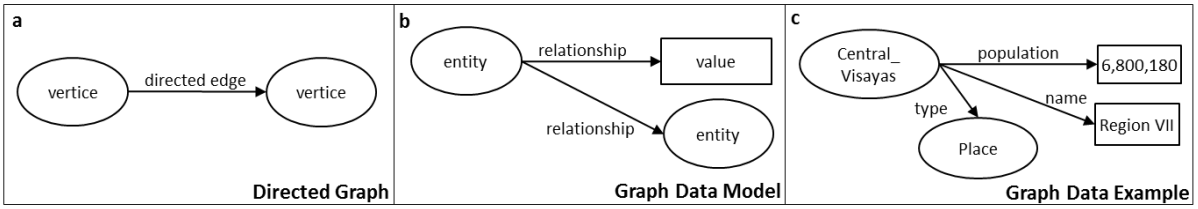


Figure 5: Graph data representations

Since the earliest use of the graph concepts, it has evolved to meet new demands and has been applied in several domains. In the information system domain, one of the most important demand is to express data semantics. Hence, some works have added characteristics to the basic concept of graph, creating derivations, such as property graph, semantic graph, conceptual graph, RDF graph and more. Some of these concepts were correlated and illustrated by Rodriguez and Neubauer (2010) (see Figure 6).

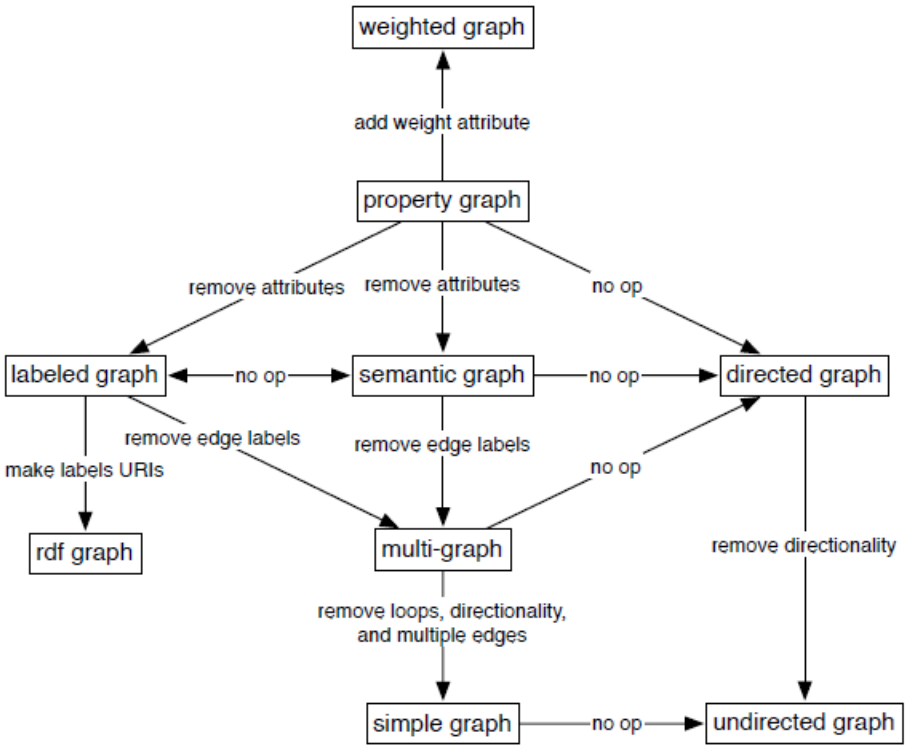


Figure 6: Graph type morphisms (RODRIGUEZ; NEUBAUER, 2010)

According to Rodriguez and Neubauer (2010), the list is not complete nor are the terms generally accepted in all domains. Many of these structures have been proposed or reused in different domains and under different names. The important point is that there are numerous graph types and, consequently, there are systems and algorithms that exist to manage and process them.

On that basis, focusing on both semantic and structure conflict resolution, some approaches improve the information integration task using conceptual and semantic graphs to manage heterogeneous information fusion (LAUDY et al., 2007; NICOLLE; CRUZ, 2009). Laudy et al. (2007) proposed an approach and a framework dedicated to high-level and heterogeneous information fusion based on conceptual graphs. Nicolle and Cruz (2009) proposed an approach based on the use of semantic adaptive graphs. The adaptive feature of the proposal makes it possible to manage two specific aspects related to information integration: the adaptation of information according to the user's access rights and the lifecycle of the integrated information. Mahfoudh et al. (2013) proposed the use of the graph grammars to formalize and manage ontologies evolution. The objective is to present an a priori approach of inconsistency resolutions to adapt the ontologies and preserve their

consistency. A framework composed of different graph rewriting rules is proposed and presented using an algebraic graph grammar tool. Taheriyan et al. (2013) proposed an approach for semantic enrichment based on a graph to hypothesize a rich semantic description of a new target source from a set of known sources that have been modeled over the same domain ontology.

Following some of the presented works, the semantic representation using a labelled graph is performed using the RDF standard and has been used as a powerful tool for information integration, as explored in the next section.

3.1 RDF Triple Representation

The vertices and directed edges of the graph data model can be implemented using the Resource Description Framework (RDF). The vertices represent the information resource, and the edges represent interlinks between them. An RDF Graph consists of a set of triple statements in the <subject, predicate, object> form (see Figure 7a). The subject identifies the resource to which the statement refers.

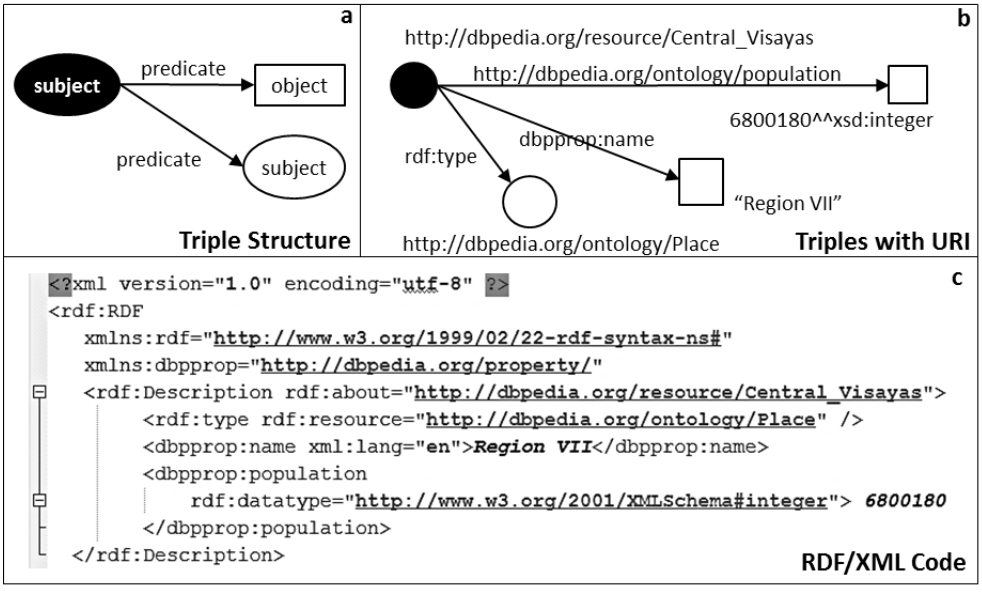


Figure 7: Graph representation of triple structure with corresponding URIs and RDF/XML code

As presented in Figure 7, the resource is identified by a Uniform Resource Identifier (URI), composed of a Uniform Resource Locator (URL) and a name, e.g., `http://dbpedia.org/resource/Central_Visayas`. The predicate corresponds to a property or relationship for this resource, e.g., `http://dbpedia.org/ontology`

/population or http://www.w3.org/1999/02/22-rdf-syntax-ns#type, respectively. Similarly, the object represents a (literal) value or a resource associated with the subject through a predicate, e.g., 6800180^^xsd:integer or http://dbpedia.org/ontology/Place respectively. The RDF statements can be serialized as RDF/XML code (see Figure 7c).

In order to provide a more precise definition of RDF nodes, triples and graph, some authors have proposed a formalization as follows (YONGTAO et al. 2013; MICHELFEIT et al., 2014).

RDF nodes: Let U, B, and L be sets of all URI references, blank nodes and RDF literals, respectively. Sets U, B, and L are pairwise disjoint. An RDF node is an element of their union.

$$N = U \cup B \cup L.$$

RDF triples: An RDF triple is a statement expressing that a resource has a property with a certain value. Formally, the set of all triples is:

$$Triples = (U \cup B) \times U \times (U \cup B \cup L).$$

RDF graph: A subset G of Triples can be represented as a directed labelled graph and referred to as an RDF Graph.

RDF is a standard specification of the World Wide Web Consortium (W3C) originally designed as a metadata model which enables data interchange on the Web. According to its specification, “RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed.” RDF features are suitable for the requirements of the solution approach proposed in this thesis. Besides, RDF specifications became a standard for data publication on the Web.

3.2 Web of Data

Based on graph structure, represented by RDF statements and supported by the W3C, the Web of Data involves simple principles of the Semantic Web to interlink and annotate data, under an open source license, reusing vocabularies or schemas (BERNERS-LEE, 2006). The graph data model is based on the Web of Data and has been taken as an alternative to solve conflicts about the structure and the semantic levels of data. Also known as Linked Open Data (LOD), it is a powerful environment, which can be used, for example, as

a worldwide source of information (BIZER et al., 2009). At the core of the LOD initiatives resides the idea of interconnecting fine-grained information resources not originally associated, leveraging on the Web infrastructure by using HTTP, URIs, and RDF as a data representation framework.

3.2.1 Interlinking Main Functions

The fundamental principle of LOD is about data interlinking. On the Web of Data, the links are primarily intended to aggregate value to data. They can be used to describe, associate, and infer new knowledge, besides supporting navigation and exploration. The interlinking can be used as an important semantic enrichment mechanism: (i) to describe information resources through an association to other information resources that corresponds to a new attribute; and (ii) to associate information using new interlinks on the triple, graph, or dataset level. Descriptions can contextualize the data, improve semantics, add new information, reveal the source of the data, and even add simple values as its type.

The interlinking can be used to describe information resources. The description of a resource can be enriched through an interlink with another information resource that describes a new attribute. The example in Figure 8 describes the Ardbeg Whisky showing the region of its origin.

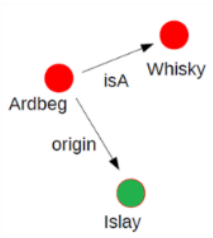


Figure 8: Sample of a resource description using a link

The interlinking can be used to associate information. Different information domains can be associated using new interlinks on the triple, graph, or dataset level. In the example illustrated in Figure 9, the Whisky domain is integrated with the geographic area through the origin place of each type of Whisky.

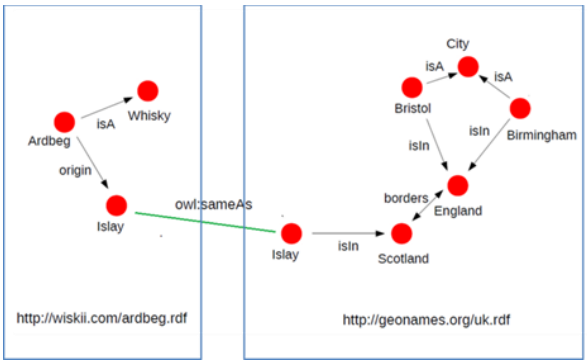


Figure 9: An example of integration of domains through an interlink

The interlinking can be used to discover or infer new information. New information can be discovered by navigating through the links from a resource to another, as well as new connections can be inferred. The dashed arrows of Figure 10 represent the inferred links. Relations with wider coverage may emerge, such as: *DerivedFrom*, *SnapshotOf*, *mayHaveBeenDerivedFrom*, *HighlyCorrelatedWith*.

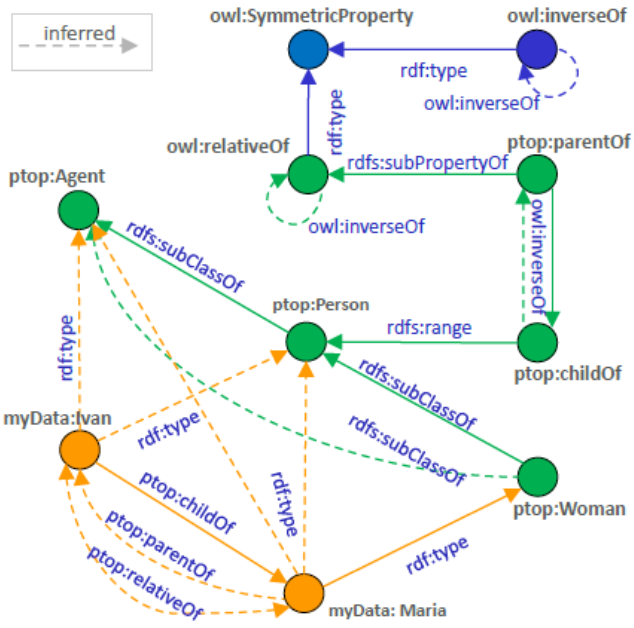


Figure 10: Example of inference to the discovery of new information through interlinks

3.2.2 Data Quality and Provenance on the Web of Data

One of the main concerns while consuming data published on the Web is trustworthiness, one criteria of data quality. Data quality is a complex concept considering its multidimensionality and relativity (AMARAL; CAMPOS, 2004). It has different meanings for different people, so it is necessary to define quality characteristics to describe how the

assessment will be made. In databases, two types of assessments can be made: a quantitative (or objective) and a qualitative (or subjective) assessment. On a quantitative evaluation, objective indicators are used and they are usually accomplished by automated tools (software). A qualitative assessment of quality depends on the observers' point of view and on the data being measured. Some data quality criteria are: reliability, credibility, consistency, completeness, relevance, interpretability, ease of understanding, timeliness, reputation, relevance, accuracy, syntactic, and semantic accuracy (AMARAL; CAMPOS, 2004).

Studies indicate that the approach to assessing data quality on the Web of Data analysis is through the origin of information (HARTIG, 2009). Provenance has been investigated in the areas of scientific workflow and database (CRUZ et al., 2009). Provenance information about an item of data represents the history of the item, starting with its creation, including information about its origins. Hartig (2009) proposes that the source of data items on the Web of Data includes information about their access and creation. On access, information is obtained about the Web publisher and provider that stored the data. On creation, descriptions are obtained about the implementation of the process of creating the data.

Data quality in the Web of Data is the subject of several studies (DIVIDINO et al., 2009; FREITAS et al., 2012; MENDES et al., 2012; ZHAO; HARTIG, 2012; MENDONÇA et al., 2013; MICHELFEIT et al., 2014) and represents an important aspect of the architecture proposed in this thesis.

3.3 Conventional Database x Triple Stores

Triple stores are a type of graph databases designed to store triples and are implemented with features that enable Web data access through an endpoint. Drawing a parallel between traditional databases and triple stores, it is possible to observe functional differences revealing strength and weaknesses of these data persistence environments. Table 2 lists some database features and relates it with the way these two environments handle these features. Notice that, the triple store approach presents some advantages with regard to complex solution requirements, particularly those concerned with schema creation and information requirements. A schema in relational databases defines tables and explicitly lists attributes each record in the table must have. A schema also defines additional integrity

constraints. In a triple store, the schema has a different function—to add value by providing an additional description of the information resource with its type.

Table 2: Conventional database x triple stores

Feature	Conventional Database	Triple Store
Schema Creation	Before instances	After, before or does not exists
Schema Implementation	On data dictionary	On triples
Schema Function	Structure and add value	Add value
Adaptation to new requirements	Weak	Strong
Data Quality	High	Defined by metrics
Entity Identification	Through tables	Through context of use
Relationship	Only relations between tables are explicit	All relations are explicit
Integration	Internal: high External: low	Depends on requirements and effort employed
Semantic expressiveness	On conceptual model	On relations
Queries	Structured, mainly local	Based on keywords, mainly distributed
Queries results	Precise and homogeneous	Ranked and heterogeneous
Management Support	Advanced and stable	In progress
Inference	Through mining algorithms	Through links navigation
Create, Read, Update and Delete (CRUD)	All covered	Create and read, evolving to update and delete
Atomicity, Consistency, Isolation, Durability (ACID) Properties	Guaranteed	Not guaranteed

3.4 Heterogeneous Data Integration Supported by Linked Open Data

The use of Semantic Web technologies for data integration purposes has been explored by industry applications (GRAUBE et al., 2011; AARNIO; SEILONEN, 2013; NEČASKÝ et al., 2014) confirming what was stated by Breslin et al. (2010): “By mapping data contained in heterogeneous applications to common representation layers using RDF(S)/OWL, they unify heterogeneous data structures in a meaningful way.” This idea was inspired by the RDF[1] Bus architecture, initially proposed by Berners-Lee (2005) and detailed by Passant (2010). It has also been used as an approach for heterogeneous data integration by

[1] http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

"translating data sources to RDF, using common ontologies for representing their semantic, then enabling data manipulation using SPARQL[2] over HTTP." These technologies have been applied to provide an integrated view of situations in the dynamic environment of emergency management (CORDEIRO et al., 2011b; SCHULZ et al., 2012a; SCHULZ et al., 2012b; MIJOVIC et al., 2013; CORDEIRO et al., 2014a). This researche has produced strong evidence that linking features of LOD address many problems of dynamic environments.

On the Web of Data, integration occurs when different domains are interlinked by semantic relations, when data is interlinked to conceptual reference frames also published in LOD format, and annotated on widely used vocabularies and ontologies. As data is associated with other information, an integrated view can be built to support decision making at different levels and scope. Furthermore, as data are analyzed, more data can be retrieved by navigating through the links and enabling the exploration of the domain. The the interlinking process provides a powerful mechanism for heterogeneous data integration as used in approaches and supporting tools.

Notice that the concept of integration in the Web of Data may differ from the concept of integration based on a relational database. In the Web of Data, two or more data sources can be integrated through the creation of a new dataset only composed by the links that kept the sources as they are. In relational databases, two or more data sources can also be integrated through the creation of a new dataset, or a view of them, but it is composed by the data from the sources eliminating redundancies and inconsistencies. This procedure in the Web of Data is called fusion, where multiple records representing the same real-world object are fused into a single, consistent, and clean representation. It can be performed by solving conflicts between object values defining rules, such as maintaining the max or min value, or calculating and maintaining the medium (BLEIHOLDER; NAUMANN, 2008; MENDES et al., 2012). In this thesis, the main proposition of integration is based on the Web of Data, thus it refers to create links, even though fusion techniques can be employed.

Jain et al. (2010; 2011) proposed BLOOMS, an approach for finding schema-level links between LOD datasets similarly to ontology alignment. It is based on the idea of bootstrapping information already present on the LOD cloud. Araujo et al. (2010) presented

[2] http://www.w3.org/TR/rdf-sparql-query/

Fusion, a framework that simplifies the definition of mappings by providing a visual user interface that integrates the exploratory process and the mapping process.

The Linked Data Integration Framework (LDIF), proposed by Schultz et al. (2011), was one of the first frameworks used within Linked Data applications to translate heterogeneous data into a clean local target representation while keeping track of data provenance. In parallel, the R2RML was recommended by W3C as a language for expressing customized mappings from relational databases to RDF datasets. Extending it, Dimou et al. (2014) proposed RML, a generic language for integrated RDF Mappings of Heterogeneous Data. Broadening R2RML scope, the language becomes source-agnostic and extensible, while facilitating the definition of mappings of multiple heterogeneous sources. This leads to higher integrity within datasets and richer interlinking among resources.

Focusing on implementations support, Groth et al. (2014) discussed how Application Programming Interfaces (API) can extend the classical Linked Data application architecture to facilitate data integration. Araujo (2014) proposed an approach, also based on the Semantic Web, for data integration over distributed and heterogeneous data endpoints. The work describes an architecture for instance matching that takes into account the particularities of this heterogeneous and distributed setting. With regard to quality assessment of integrated data, several works have proposed collecting provenance data of the interlinking process (HARTIG, 2009; MARJIT et al., 2012; MENDES et al., 2012; ZHAO; HARTIG, 2012; MICHELFEIT et al., 2014).

Besides architectures, languages, frameworks and APIs supporting heterogeneous data integration based on Linked Open Data, some interlinking algorithms have been developed to discover links between datasets (FERRARA et al., 2011; SCHARFFE; EUZENAT, 2011; WÖLGER, 2011) and to fuse data sources (LAUDY et al., 2007; MENDES et al., 2012; MICHELFEIT et al., 2014) based on conflict resolution.

3.5 Supporting Tools, Frameworks and Architectures for Linking Data

Some transformation steps, such as cleaning, conversion, and conforming, precede the data interlinking process. Other steps can be added, for example, annotating the data according to conceptual models and increasing the semantic expressivity. Focusing on providing a tool to support this data transformation and linking steps, some works focus on

developing stable and friendly development environments. The tools are usually composed of a set of components and modules that are continuously evolving. Some examples are: Linked Data Integration Framework (LDIF/R2R/Silk/Sieve), Extract, Transform and Load framework for LOD (ETL4LOD/ETL4LinkedProv), Open Data Clean Store (ODCleanStore/ODCS-FusionTool/UnifiedViews), and a Service Oriented Architecture for Linked Data Integration using Syntactic, Structural and Semantic Mediation for Service Composition (DMaaS) approach.

The Linked Data Integration Framework (LDIF) (Figure 11), initially proposed by Schultz et al. (2011), offers a modular architecture to support a wide range of applications in producing a homogenized view over heterogeneous data originating from diverse sources. Data sets are imported into LDIF through *Web Data Access Modules*, including RDF dump import, Web of Data crawler, and SPARQL endpoint import. The *Data Translation* module employs the R2R Framework, proposed by Bizer and Schultz (2010), to translate Web data that is represented using terms from different vocabularies into a single target vocabulary. Vocabulary mappings are expressed using the R2R Mapping Language[3].

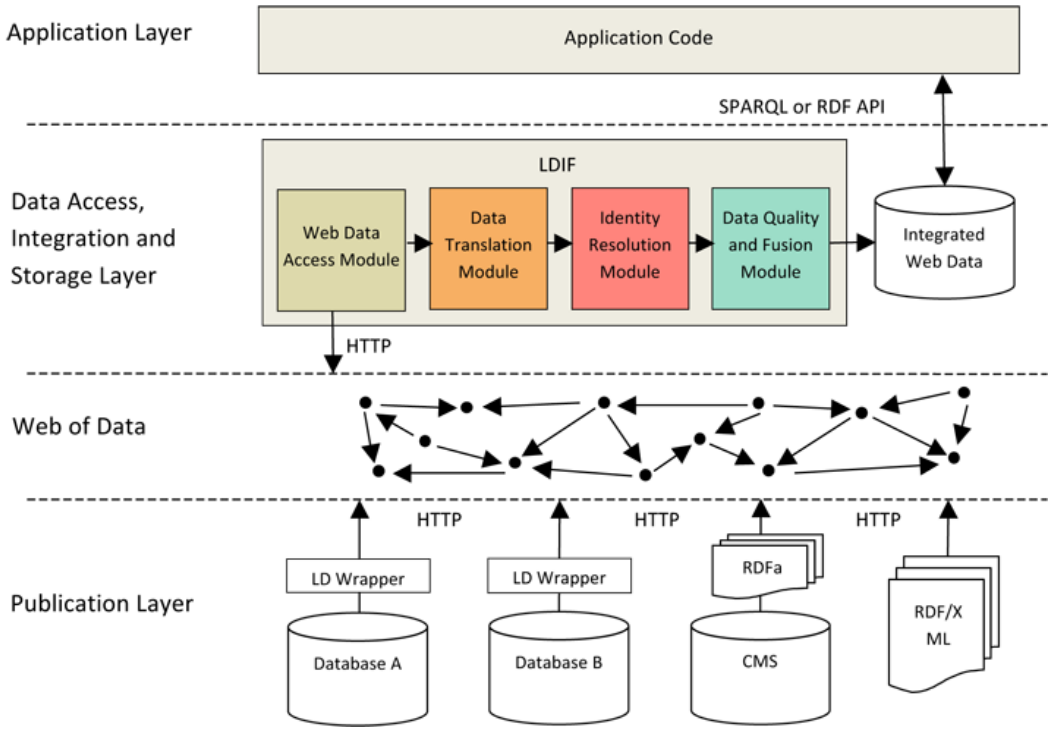


Figure 11: Linked Data Integration Framework (LDIF) architecture (SCHULTZ et al., 2011)

[3] http://wifo5-03.informatik.uni-mannheim.de/bizer/r2r/spec/

The *Identity Resolution* module employs the Silk Link Discovery Framework[4], proposed by Volz et al. (2009), to find different URIs that are used within different data sources to identify the same real-world entity. For each set of duplicates identified by Silk, LDIF replaces all URI aliases with a single target URI within the output data. The Data Quality Assessment and Fusion module employs Sieve, proposed by Mendes et al. (2012), to provide data quality evaluation and cleansing.

Similar to LDIF, ODCleanStore is a server application for integration and management of Linked Data. Initially proposed by Knap et al. (2012), it accepts Linked Data as RDF, processes them in a customizable pipeline of data processing units and saves the result to a data store. Users are provided with integrated views on the processed data that are generated on demand by the data fusion component, called ODCS-FusionTool, later proposed by Michelfeit et al. (2014) (Figure 12). Thereafter, also based on ETL paradigm, the UnifiedViews[5], proposed by Knap et al. (2014), is a framework that allows users to define, execute, monitor, debug, schedule, and share ETL data processing tasks, which may employ custom plugins created by users. An important feature of UnifiedViews is that it natively supports RDF data and ontologies.

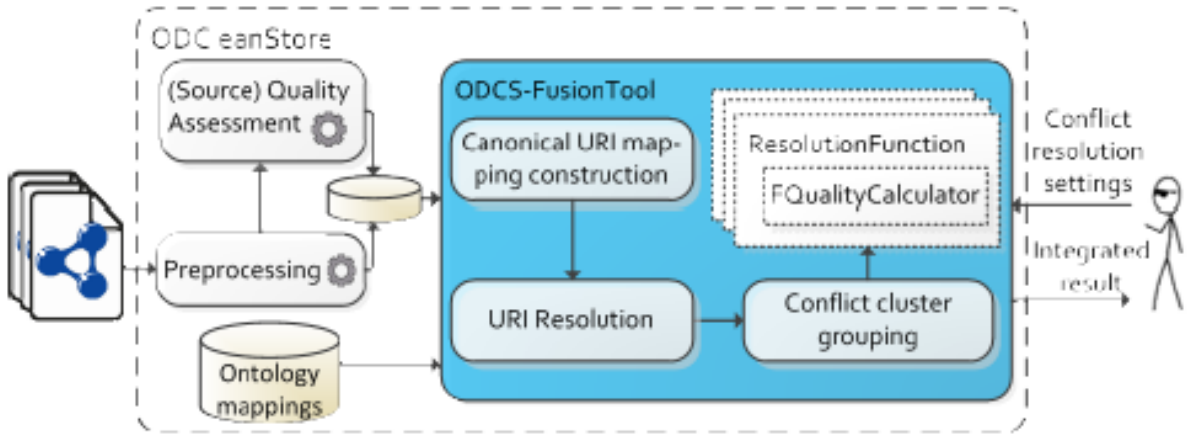


Figure 12: ODCleanStore and ODCS-FusionTool framework (MICHELFEIT et al., 2014)

[4] http://silk-framework.com/

[5] http://www.unifiedviews.eu/

Developing a broader set of steps for data integration, Cordeiro et al. (2011a) implemented the ETL4LOD, which adds a set of extensions to the ETL workflow engine called Pentaho[6]. The plugins support the data transformations required to pre-process, to integrate, and to interlink heterogeneous data. Thereafter, Mendonça et al. (2013) proposed the Linked Open Provenance (LOP) (Figure 13), and used the same platform to support the collection of prospective and the retrospective provenance data of the process. The implemented set of extensions is called ETL4LinkedProv[7] (MENDONÇA, 2013). With this tool, it is possible to interlink the provenance data with the corresponding transformed data, enabling sound data quality assessments.

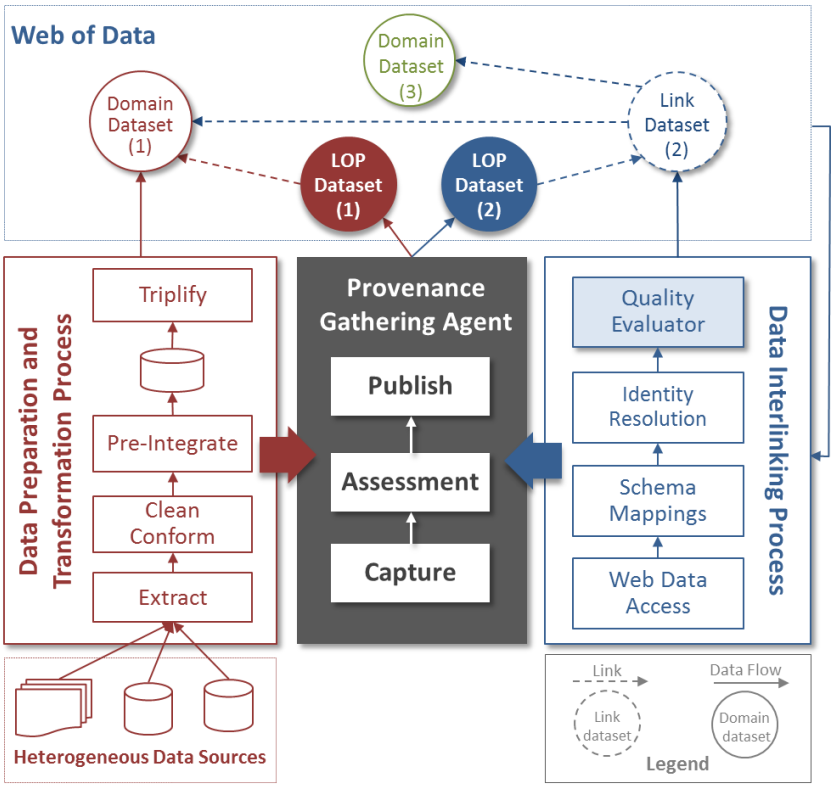


Figure 13: Linked Open Provenance (LOP) architecture overview (MENDONCA et al., 2013)

Table 3 summarizes some of the main characteristics of supporting tools, frameworks, and architecture for linking data. Notice that each one has its strengths and weaknesses relative to its objectives and application case. In this thesis, to support the solution for the open issues explored in Section 2.3, the ETL4LOD architecture was extended to allow the reuse of

[6] http://www.pentaho.com/product/data-integration

[7] http://greco.ppgi.ufrj.br/lodbr/index.php/principal/etl4linkedprov

interlinking tools, such as Silk, and to natively process RDF Graph as the UnifiedView. Details of the architecture and the supporting development framework are described in Section 4.3.

Table 3: Tools for Linking Data

Approach	Linked Data Integration Framework	Extract, Transform and Load framework for LOD	Open Data Clean Stores
Tools	LDIF/R2R/Silk/Sieve	ETL4LOD/ETL4Linked Prov	ODCleanStore/ODCS-FusionTool/UnifiedView
ETL Features Covered	Basic string transformation	Several, natively from ETL tools	SPARQL functions
Processed Grain	RDF Triples	Tuples and RDF Triples	RDF Triples and RDF Graph
Interfaces Supported	SPARQL endpoint, file system	SPARQL endpoint, file system, Relational Databases	SPARQL endpoint, file system
Platform	Developed from scratch	Pentaho Data Integration	Developed from scratch
Provenance	No	Yes	No

More recently, De Vettor et al. (2014) have been developing an independent data source and service-oriented architecture that offers the mechanisms to detect and resolve data heterogeneity issues, and provides tools to aggregate and process information in order to generate smart data from diverse Web data sources (Figure 14).

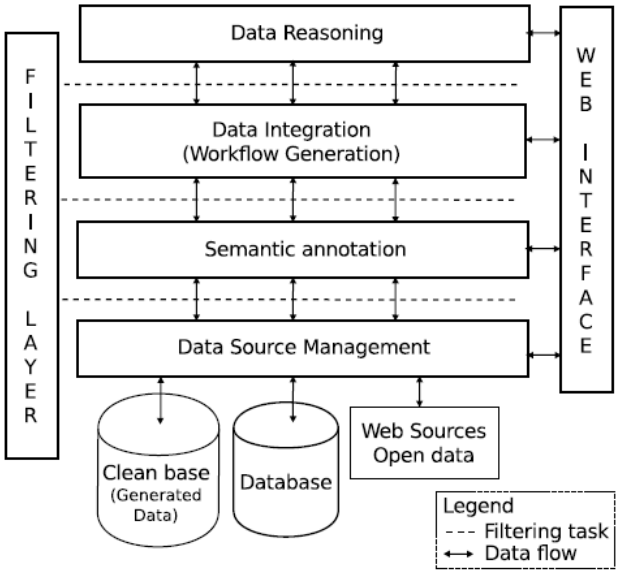


Figure 14: Service oriented architecture for linked data integration (DE VETTOR et al., 2014)

The authors define smart data as significant, semantically explicit data, ready to be useful to fulfill the stakeholders' objectives. Each operation, as well as smart data access, is available through a linked data service. A linked data service is a Web resource identified via an URI and accessible with HTTP verbs. In the data integration layer, to detect possible heterogeneity problems, could they be syntactic or structural, and to reconcile them, the architecture uses the DMaaS approach, proposed by Mrissa et al. (2013), an automated solution for resolving data heterogeneity problems between semantically described data, using a decentralized (peer to peer) repository of mediation services (a Web services dedicated to data conversion). The DMaaS approach classifies data heterogeneity issues according to the syntactic, structural, and semantic levels, and provides adapted mediation along these levels. The approach sets up the automatic conflict detection mechanism, which analyses input metadata, and intercepts data responses to perform reconciliation (SELLAMI et al., 2014).

Despite several works that have been published addressing the integration of heterogeneous data using LOD and the broader issues covered by tools with sound results achieved, there is still a lack of adequate facilities that interlink data from unpredicted sources concerning semantic expressivity level heterogeneity. This thesis addresses these open issues, also based on LOD, as described in the following chapters.

4 aDApTA: Integration of Emergent Information in Complex Environments

Different from complex systems, simple or complicated systems can be characterized by the continuous flow of interactions, i.e., without disruptions, among the elements that compose an environment. In these systems, the interactions have been planned, follow a social code or norm, or act as specified. However, we argue that, in complex systems, an element can change its behavior in an unpredicted way, causing the interruption of the interaction flow. This interruption can be viewed as a conflict between two elements that cannot communicate with each other any longer. This conflict has to be resolved in order to establish a new flow of interactions. The conflict resolution implies changing the other elements' behavior. This change can be called adaptation. Thus, in short, adaptation can be seen as a way of resolving conflicts between the elements of an environment by changing components' behavior in order to restore a dynamic equilibrium of the system flow. An element can be, for example, a data resource in an information system.

Traditionally, a data resource can only be processed by an information system fulfilling its constraints. If a resource with a different file format or implicit meaning emerges in an unpredicted way, the system may not be able to process it, causing conflicts. In this scenario, how can a system detect and resolve information heterogeneity conflicts? To answer this question, this work proposes an approach, named aDApTA, for dealing with structural and semantic expressivity conflicts in order to enable new flows in the information systems.

The approach is composed of a set of systematically organized activities in a process, as illustrated in Figure 15. The activities are grouped into two sections: the first handles the adaptive integration of the unpredicted information, concerning structural and semantic expressivity level conflicts. The structural conflict is solved converting the incoming data to RDF Graph, and the semantic expressivity conflict is solved annotating the information with terms of common vocabulary and concepts of domain ontologies. Once the conflicts are solved, the interlinking is performed using an appropriate tool chosen according to the

semantic expressivity level of the information. The result is the integration of the incoming data, representing the current knowledge, and the domain information base, representing the previous knowledge, composing the combined knowledge, as explored in Figure 2 of Section 2.1.

The second group of activities of aDApTA handles the provenance collection and its interlinking with the interlinked domain data. The collection is performed in parallel to the adaptive integration activities. Only after that, the provenance is published using terms and concepts from common provenance vocabulary and ontologies, and then it is interlinked with the domain data allowing the joint exploration. The Business Process Model and Notation (BPMN), standardized by OMG (2011), is used to represent the aDApTA process. However, an extension was needed to depict the provenance monitoring. To address that, a double line circle with the letter P was used.

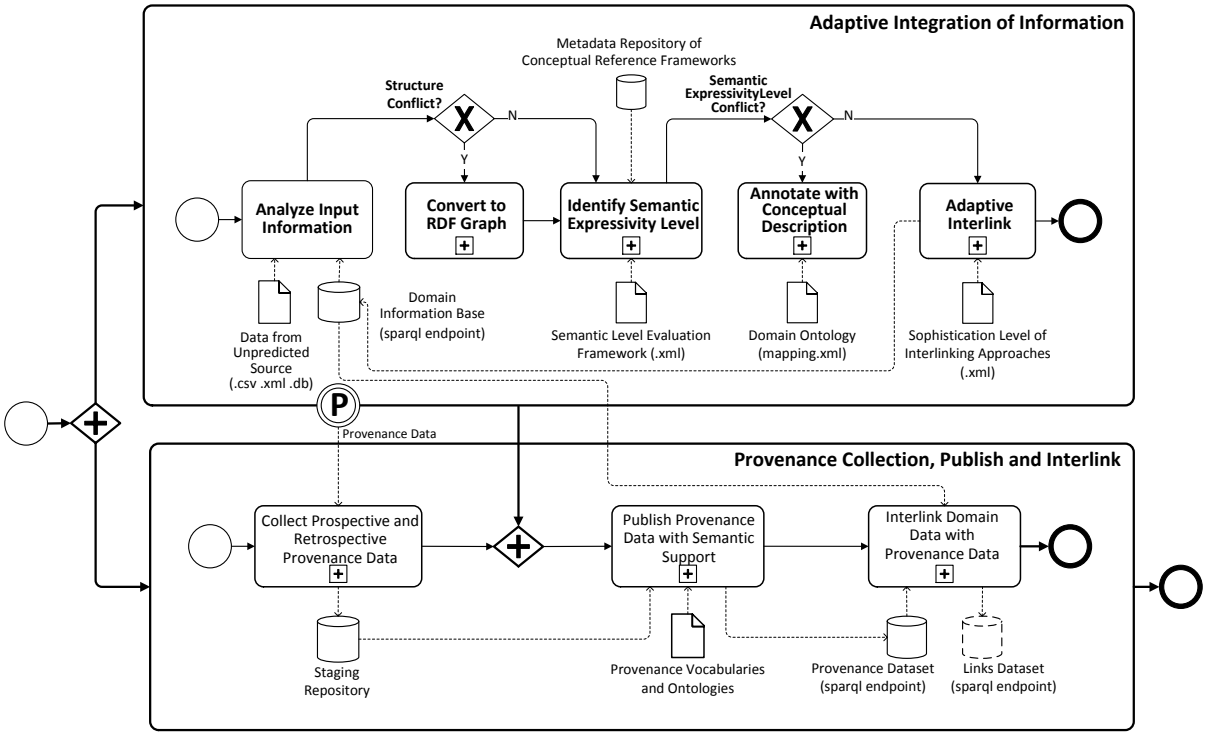


Figure 15: Adaptive approach for information integration with provenance

4.1 Adaptive Integration of Information

The flow of the first section starts with two basic activities: *Analyze Input Information* and *Convert to RDF Graph*. The former analyzes the incoming data through a technique, called data profiling, used to identify the structure and to become aware of the content of a source in order to check anomalies before building a cleaning and conforming process (KIMBALL et al. 2013). If there is a structure conflict, i.e., if it has a different format from the domain information base, the second activity of the approach converts the data to an RDF Graph. This process is also known as triplification and it is implemented through a set of steps in a workflow. Once implemented, the triplification runs automatically. Transformation tasks are also performed in this activity to solve the anomalies detected in the previous activity. The other three activities of the *Adaptive Integration of Information* section are described below. As they represent the core of the aDApTA approach, each activity is detailed in one section.

4.1.1 Identify Semantic Level

After the conversion, the RDF Graph semantic expressivity level is assessed. This is a central activity of aDApTA in enabling the adaptation of the interlinking approach. The activity is performed by using a *Semantic Level Evaluation Framework to* identify the description associated with a resource and to assign a level to it. The framework is configured through an XML file with a value and a label used to represent each level. Furthermore, an evaluation rule is set. The rule will be executed by the implementation code of the ETL4LOD-Graph, as detailed in Section 4.3.2 (ETL4LOD-Graph Framework). Figure 16 shows a template example of the framework configuration XML file.

```
<SemanticLevelFramework>
    <Level id="1">
      <Rule></Rule>
      <Value></Value>
      <Description> </Description>
   </Level>
<SemanticLevelFramework>
```

Figure 16: Template example of the *Semantic Level Evaluation Framework* configuration XML file

Applying the framework configuration to the examples illustrated in Figure 17 and Figure 18, some possible semantic expressivity levels are:

- LOW: a resource described by a literal;
- MEDIUM: a resource described by other resource;
- HIGH: a resource or a property described by a prefix, which represents a vocabulary or an ontology; and
- ADVANCED: a resource described by an ontology concept modeled based on a foundational ontology[8].

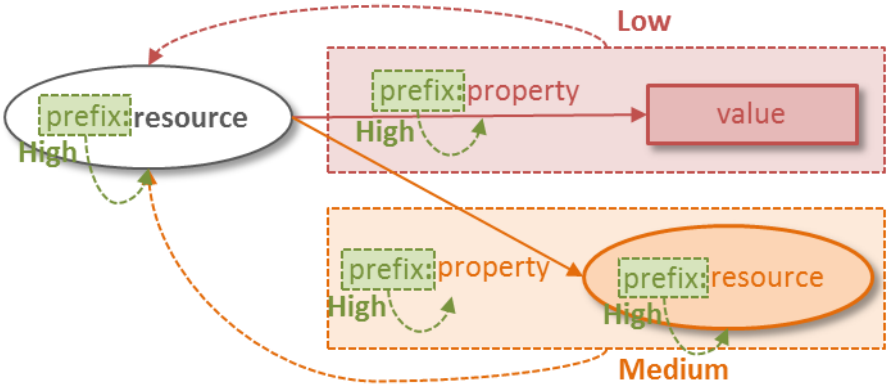


Figure 17: Diagram of some semantic expressivity levels of an information resource

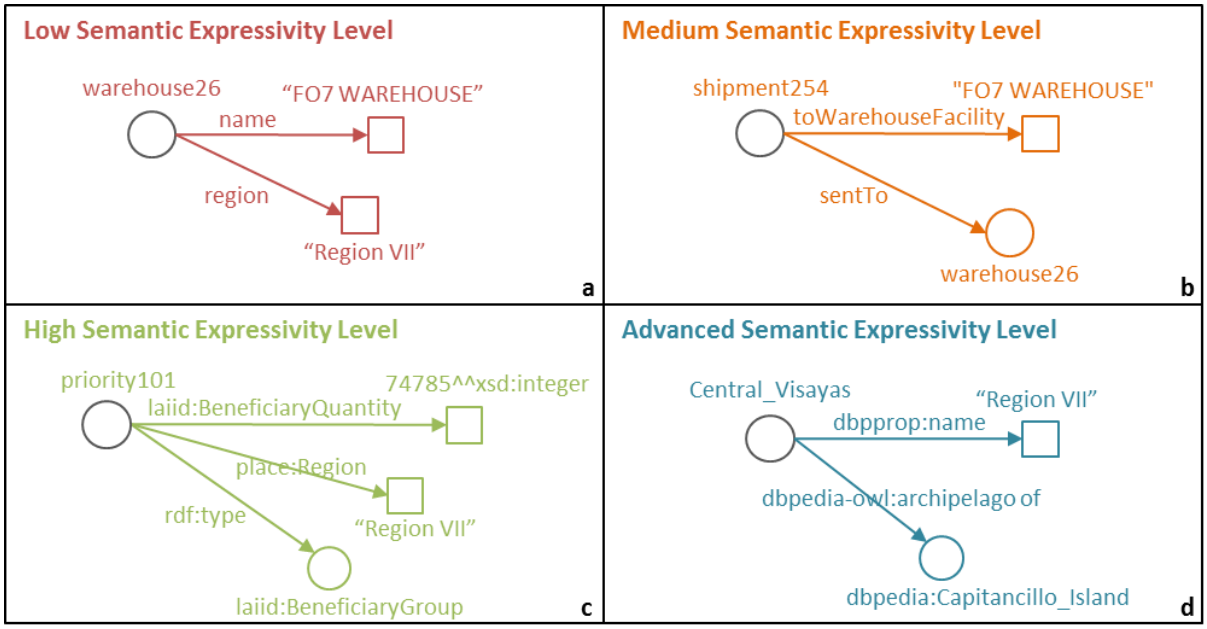


Figure 18: Examples of graph data with different semantic expressivity levels

[8] A foundational ontology is composed of well-founded high level concepts which are used to improve the semantic expressivity of domain concepts (see www.inf.ufes.br/~gguizzardi/BalticDB&IS.pdf)

The semantic expressivity level can be assessed, for example, by checking if the prefix used to describe the subject and predicates of an information resource represents the concepts of a vocabulary or those of an ontology. If it represents a vocabulary, then the subject item has lower semantic expressivity than the one described by an ontology. For example, the information resource *priority101* has lower semantic expressivity level than *Central_Visayas*, because it is basically described by concepts of domain ontologies (Place, RDF and LaiiD) and the latter is described by concepts of an upper ontology (dbpedia-owl).

Notice that these levels belong to a simplified framework with a basic semantic structure of an information resource. More levels can be added. For example, an information resource described with 20 literal properties has a lower semantic expressivity level than a resource described with 10 properties mainly annotated with ontology concepts. Hence, the more detailed the framework is, the more accurate results are expected from the adaptive approach for information integration, i.e., more expressive interlinks. However, more data processing and time are also required. The volume of data and the time requirement have to be evaluated to determine the framework detailed level, together with requirements of expressiveness precision of the application case and the integration approaches available.

The conceptual framework, depicted in Figure 19a, was used as a reference to create a vocabulary, called Semantic Expressiveness Level Stamp, abbreviated as `sstamp`.

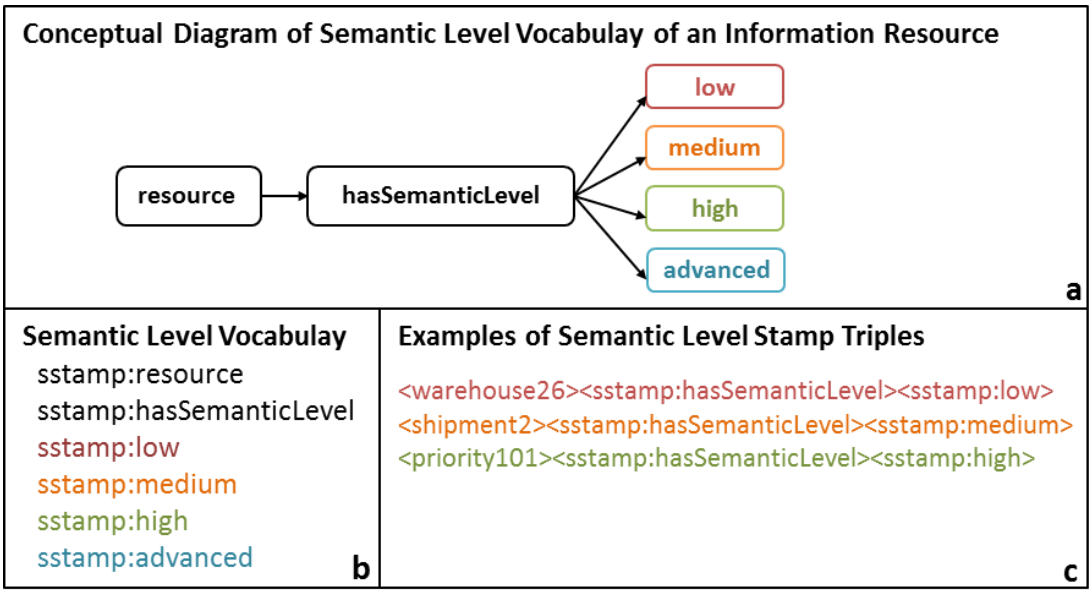


Figure 19: Simplified semantic interlinking level vocabulary and examples

As illustrated in Figure 19b and Figure 19c, the `sstamp` vocabulary is used to mark the triples with the identified semantic expressivity level, which supports the selection of the appropriate interlinking approach. This step is accomplished using a program code which creates new stamp triples according to the designed framework, e.g., `<resource><sstamp:hasSemanticLevel><sstamp:low>`. The implementation and application of this conceptual framework is detailed in Sections 4.3.1 (Architecture Components) and 4.3.2 (ETL Workflow of the Adaptive Interlinking), respectively.

Algorithm 1 is the pseudo-code of this assessment process that uses the presented evaluation framework. The input is an RDF Graph and the output is a stamp triple. A *stamp triple* is composed of the subject being evaluated, the property `hasSemanticLevel`, and a resource that represents the semantic expressivity level.

Algorithm 1: Semantic Expressivity Level Assessement

```
Input: RDFGraph g[List Triple t[Subject s, Predicate p, Object o]]
Output: Triple stamp[Subject s, Predicate o, Object o]
initialization;
i ← 1;
valueLevel ← 0;
assessedValueLevel ← 1;
semanticLevel ← "low";
forall the Triple t in g do
    prefixS ← g.t(i).[s].getPrefix();
    prefixP ← g.t(i).[p].getPrefix();
    prefixO ← g.t(i).[o].getPrefix();
    if g.t(i).[o] is Literal() then
        semanticLevel ← "low";
        assessedValueLevel ← 1;
    if isVocabulary(prefixS) or isVocabulary(prefixP) or isVocabulary(prefixO) then
        semanticLevel ← "medium";
        assessedValueLevel ← 2;
    if isOntology(prefixS) or isOntology(prefixP) or isOntology(prefixO) then
        semanticLevel ← "high";
        assessedValueLevel ← 3;
    if valueLevel <assessedValueLevel then
        stamp ← [g.t(i).[s], "hasSemanticLevel", semanticLevel];
        valueLevel ← assessedValueLevel;
    i ← i + 1;
```

Figure 20 shows a RDF Graph marked with a *low stamp triple*, as it is only described by literals, and another RDF Graph marked with a *high stamp triple*, as it is described by other resources and concepts from ontologies.

Figure 20: Examples RDF Graphs marked with a *stamp triple*

In order to discover if the prefix represents a vocabulary or an ontology, the approach implementation uses a *Metadata Repository of Conceptual Reference Frameworks*. The Linked Open Vocabulary (LOV)[9], for example, stores some properties about the vocabularies and ontologies, such as, name, prefix, title and description. In aDApTA, using the simplified framework depicted in Figure 17, the identification is performed by two functions that search the terms "vocabulary" or "ontology" in the description property. Additionally, it is possible to recognize if the prefix represents an upper ontology by querying the tag field with "General & Upper". Currently, there are 13 ontologies tagged as a general and upper ontology. Some examples are: Descriptive Ontology for Linguistic and Cognitive Engineering + Descriptions and Situations (DOLCE+DnS) Ultralite[10], British Broadcasting Corporation (BBC) Core Concepts[11], and Upper Mapping and Binding Exchange Layer (UMBELL)[12].

[9] http://lov.okfn.org/dataset/lov

[10] http://www.ontologydesignpatterns.org/ont/dul/DUL.owl

[11] http://www.bbc.co.uk/ontologies/coreconcepts

[12] http://umbel.org/umbel

For simplification purposes, a dump of LOV repository was downloaded to feed the system. However, it would be possible to use the LOV endpoint[13] or API[14] for online access. Figure 21 illustrates the SPARQL query and Table 4 presents a subset of the corresponding result.

```
PREFIX vann:<http://purl.org/vocab/vann/>
PREFIX voaf:<http://purl.org/vocommons/voaf#>
PREFIX dc: <http://purl.org/dc/terms/>

### Vocabularies contained in LOV, their prefix, title and description
SELECT DISTINCT * {
    GRAPH <http://lov.okfn.org/dataset/lov>{
        ?vocabURI a voaf:Vocabulary.
        ?vocabURI vann:preferredNamespacePrefix ?vocabPrefix.
            ?vocabURI dc:title ?vocabTitle.
            ?vocabURI dc:description ?vocabDescription.
}} ORDER BY ?vocabPrefix
```

Figure 21: SPARQL query performed in LOV endpoint

Table 4: Subset of the result of the SPARQL query performed in LOV endpoint

vocabURI	vocabPrefix	vocabTitle	vocabDescription
http://purl.org/acco/ns	acco	"Accomodation Ontology"@en	"A **vocabulary** for the description of hotels, vacation homes, camping sites, and other accommodation offers for e-commerce"@en
http://www.rkbexplorer.com/ontologies/acm	acm	"ACM Classification Ontology"@en	"This **ontology** is a representation of The ACM Computing Classification System [1998]"@en
http://privatealpha.com/ontology/certification/1#	acrt	"Agent Certification Ontology"@en	"This document specifies a **vocabulary** for asserting the existence of official endorsements or certifications of agents, such as people and organizations."@en
http://schemas.talis.com/2005/address/schema	ad	"Address Schema"@en	"A simple **vocabulary** used to represent people address elements"@en
http://www.ontologydesignpatterns.org/ont/dul/DUL.owl	dul	"DOLCE+DnS Ultralite"@en	"The DOLCE+DnS **Ultralite ontology**. It is a simplification of some parts of the DOLCE Lite-Plus library" @en

The use of a metadata repository, such as LOV, is one of the possible approaches to discover which type of conceptual framework the prefix represents. Moreover, it is possible to have conceptual models with different levels of precision. Today, the LOV repository is available and is being used as a reference. It contains 469 vocabularies and more than 46,000 terms. The current procedure for a vocabulary registration is the suggestion

[13] http://lov.okfn.org/endpoint/lov

[14] http://lov.okfn.org/dataset/lov/api

performed by the authors who provide a few metadata properties using the LOV recommendations[15] and their own conceptualization of what a vocabulary or ontology is. However, it can evolve to more sophisticated procedures, such as, the automated recording and the input of more details about the conceptual models or vocabularies.

Despite possible limitations and imprecisions of the LOV repository, it is a useful tool to support the identification of the semantic expressivity level of the data and to further support heterogeneity conflicts resolution. The same information resource described in different ways may compromise the efficiency of integration/matching algorithms. The proposed approach aims to minimize this impact in order to improve the integration results. To do that, the semantic expressivity level of the information resource must be identified.

4.1.2 Annotate with Conceptual Description

If the semantic expressivity level of the unpredicted data source is lower than the level of the domain information base, then the semantic expressivity of the data source descriptor is augmented using conceptual reference models such as vocabularies and ontologies. This task can be accomplished in many ways, e.g., by adding a prefix to resources and properties or replacing properties with a prefix and a common concept. In addition, a prefix can be added to a resource or a property, abbreviating the namespace of a URI, representing a class or a relation of a vocabulary. Either a standard or a domain vocabulary can be used. As a result, the semantic expressivity of the information resource moves from a medium to high level.

For example, if an RDF Graph is represented with little semantic concern (Figure 18a), mainly described by literals (text strings) and property without prefixes, then it is annotated with a concept from a vocabulary or an ontology. Thus, the *Annotate with Conceptual Description* activity is performed with a mapping file that can be created manually or by an alignment approach. To perform this step, Algorithm 2 is used to annotate an RDF triple from a mapping input list.

[15] http://lov.okfn.org/Recommendations_Vocabulary_Design.pdf

Algorithm 2: Annotation of an RDF Triple

```
Input: Triple t[Subject s, Predicate p, Object o],
       List mapping[Literal from, Literal to]
Output: Triple annotatedTriple[Subject s, Predicate p, Object o]
initialization;
forall the from in mapping do
    if mapping[from]==t[s] then
        annotatedTriple[s] ← mapping[to];
    if mapping[from]==t[p] then
        annotatedTriple[p] ← mapping[to];
    if mapping[from]==t[o] then
        annotatedTriple[o] ← mapping[to];
```

Also, this task can be performed by extending existing mapping and alignment algorithms (JAIN et al. 2011; BELLENGER et al., 2013) to augment the semantic expressivity of data sources. After that, the semantic expressivity level of the new data source is assessed again to get a new stamp. Hence, the activity *Annotate with Conceptual Description* can solve the semantic expressivity level conflict enabling the new source to go further in the flow of the integration process.

Another possible conflict scenario is when the incoming data have a higher semantic expressivity level than the one already stored in the information system. In that case, a project decision must be made in order to ensure information resources with the same semantic expressivity level. The decision can be either an upgrade or a downgrade of the semantic expressivity of the information resources.

4.1.3 Adaptive Interlinking

The last activity is the *Adaptive Interlinking* that switches the interlinking approach according to the semantic expressivity level of the incoming data. The available interlinking tools use different approaches to create links. Some of them consider only the literal value of properties. Others consider the type of the information resource. There are also tools that implement algorithms that use domain ontologies to infer links between data from different sources, such as Knofouss[16] and BLOOMS[17]. In aDApTA, each interlinking tool is identified

[16] http://technologies.kmi.open.ac.uk/knofuss/

[17] http://wiki.knoesis.org/index.php/BLOOMS

according to its level of sophistication. For example, tools that use only literal values are marked with a *low stamp triple*, and tools which use ontologies to suggest links are marked with a *high stamp triple*. Similar to the semantic expressivity level framework of an information resource (see Section 4.1.1, Identify Semantic Level), a sophistication level framework was developed to support the classification of the interlinking tools. Figure 22 presents the conceptual diagram, vocabulary, and some examples of *stamp triples* of this framework.

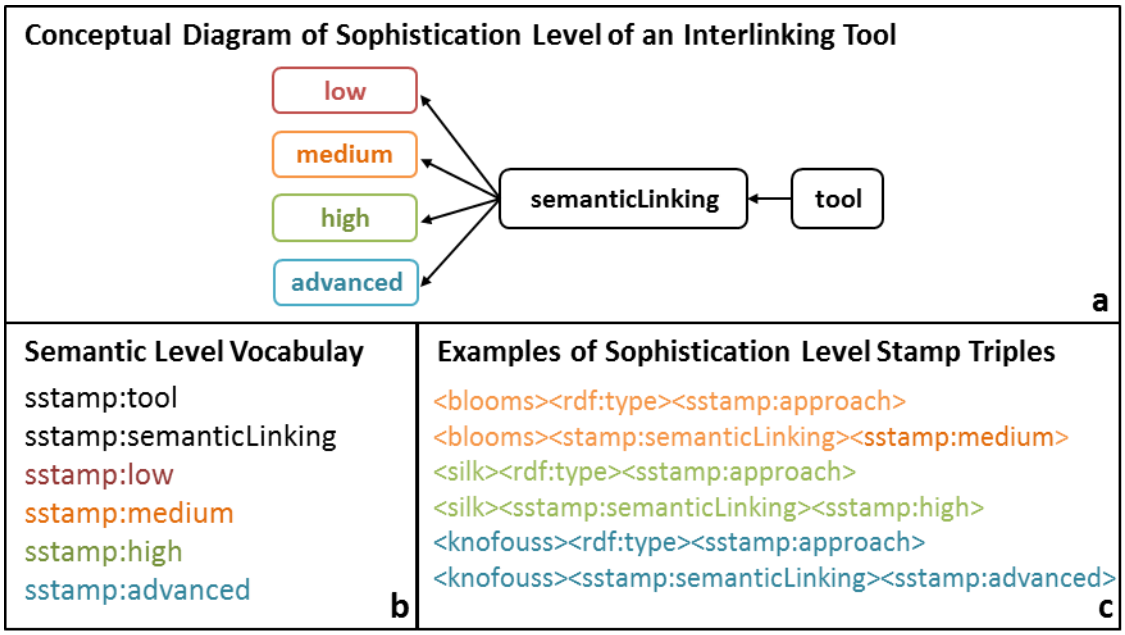


Figure 22: Sophistication levels of interlinking tools

The *stamp triples* enable the system to adapt the interlinking process by defining the corresponding and most appropriate tool for the information resources to be integrated. It is performed by selecting an interlinking tool that corresponds to the same level of the information semantic expressivity level. Different from the assessment of the semantic expressivity level of an information resource, which can be automated, the assessment of the sophistication level of a tool is performed manually. However, if there was a standard metadata repository describing the tools, similarly to LOV, the assessment could be automated too.

4.2 Provenance Collection, Publish, and Interlink

In order to support data trustworthiness, provenance data of the adaptive integration process is collected, published with semantic support, and linked to the domain data, following the approach proposed by Mendonça et al. (2013; 2014). Besides provenance of regular activities of an integration process, this approach was extended to collect provenance of the aDApTA activities: *Identify Semantic Level*, *Annotate with Conceptual Description* and *Adaptive Interlink*.

4.2.1 Collect Prospective and Retrospective Provenance Data

In parallel to the activities described in previous sections, the provenance data of the whole process is collected. The collect activity gathers prospective provenance data about the composition and specification of the process, and retrospective provenance data about the execution of the process. The scope of the collected provenance can be defined according to the application case, i.e., it is possible to collect the provenance data either from a single step or from a set of steps.

Every time an activity is performed, the provenance metadata is captured and stored in a staging repository. The monitored events are: the start and the end of the main process or any of the sub-process; the start and the end of each executed activity; the data read by a given activity. A subset of activity types can be selected to be monitored, defining the granularity of the provenance which impacts on the volume of collected data. The execution of the provenance collection activity ends when the last activity of the integration process has finished.

4.2.2 Publish Provenance Data with Semantic Support

To support the publication of provenance data semantically, the next activity uses a set of common *Provenance Vocabularies and Ontologies*. PROV-O[18] is used to represent the semantics of the process. OMPW[19] is used, as an extension of PROV-O, to distinguish the

[18] http://www.w3.org/ns/prov#

[19] http://www.opmw.org/ontology

semantics relating to prospective and retrospective provenance. Cogs[20] ontology is also used as an extension of PROV-O to represent the concepts of an ETL process (FREITAS et al., 2012). The support of these common concepts enables interoperability with the Web of Data provenance.

The provenance data captured and stored in the staging repository is annotated with these common concepts and stored in the *Provenance Dataset*, which can be accessed through a SPARQL endpoint.

4.2.3 Interlink Domain Data with Provenance Data

Finally, the collected and semantically enriched provenance data are published as LOD and interlinked with the domain data. Moreover, the retrospective provenance data are interlinked with the prospective provenance data. The generated links can be stored in the *Provenance Dataset* or separated in a *Link Dataset*. These links enable the conjunction exploration of both domain and provenance data through SPARQL queries, and supports data quality and trustworthiness assessments. Some examples are described in Section 5.5 (Decision Making Support).

4.3 Supporting Architecture

The aDApTA architecture was designed in layers grouping the elements present in the context of an environment, which supports decision making under a high flow of information. The core components are located between the data sources and the interface with the end users enabling the composition of an integrated view from heterogeneous and unpredicted data sources. In order to implement the architecture, a set of ETL steps were developed, composing a framework, called ETL4LOD-Graph, which allows the aDApTA application in any domain. Both components and development framework are detailed hereafter.

[20] http://vocab.deri.ie/cogs

4.3.1 Architecture Components

The supporting architecture (Figure 23) of aDApTA is composed of four layers. The **first layer** contains elements that represent the dynamic environment where the system is used. The main actors are the agents that contribute with their experiences on the field. For example, the *Field Agents* report their perspective of the situation, the *Decision Makers* use the information to define the actions to be performed by the *Field Agents*, and the *Information Analysts* operate the information system. Moreover, it is possible to provide open access to *Third Party* applications to use the integrated data.

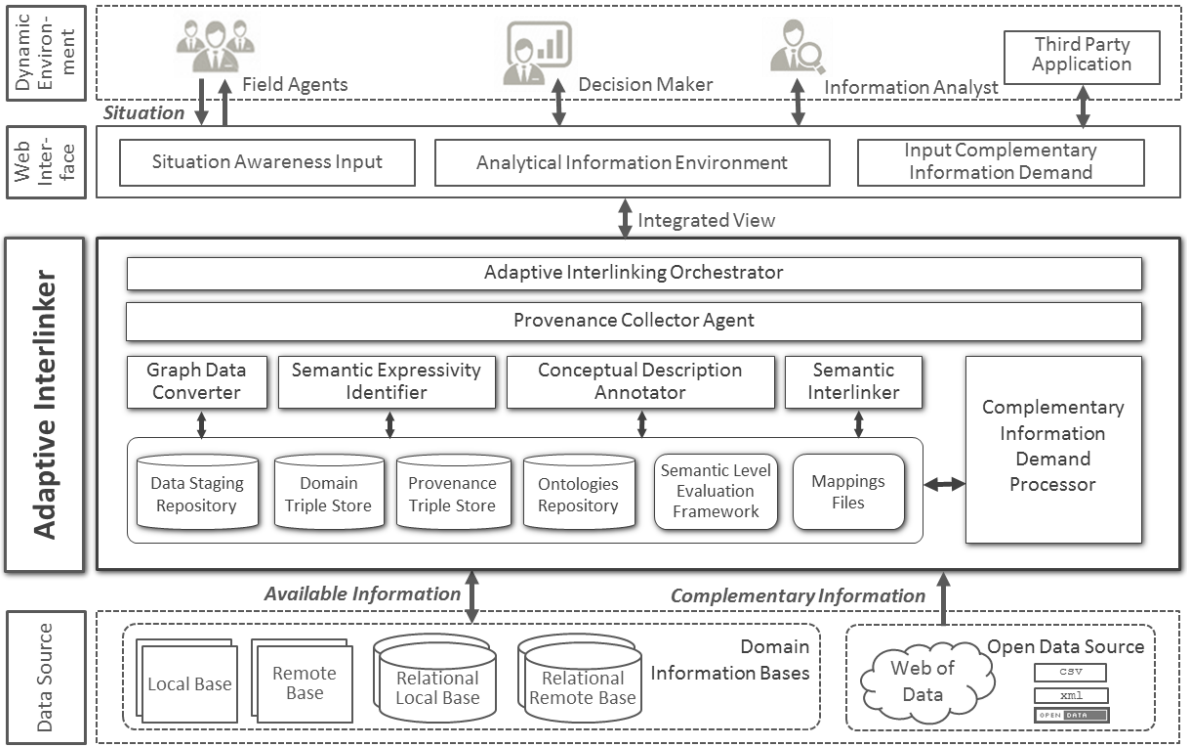


Figure 23: Architecture for adaptive data integration with provenance collection

The **second layer** is composed of Web applications acting as an interface between the integration layer and the end users. The *Situation Awareness Input* allows the input of the situation reports provided by the agents. The report can be provided in many ways as the technology evolves. For example, a field agent can use a radio to send a message to internal agents in a central room which uses a desktop application to input the information. Alternatively, the field agent can use an application available on a mobile device to report the situation directly from the field. Furthermore, it is possible to receive information about the situation from other sources, such as, sensors, social media and network, and even

images captured by drones—a technology gaining popularity recently. The current information reported by the field agents is integrated with previous information and presented to the decision maker by a dynamic and multi-perspective view. It is performed by the *Analytical Information Environment*, which provides a set of interactive graph reports showing the integrated view of the ongoing situation. Based on the analyses, the decision maker may demand new information using the interface *Input Complementary Information Demand*.

The **third layer**, called *Adaptive Interlinker*, contains the components that implement the activities and the steps of aDApTA, described in Section 0 (see Figure 15). The *Adaptive Integration Orchestrator* manages these steps. Provenance is collected by the *Provenance Collector Agent* plugged into the orchestrated steps. The *Graph Data Converter, the Semantic Expressivity Identifier, the Conceptual Description Annotator*, and the *Semantic Interlinker* implement the respective steps of the approach. Additionally, the *Complementary Information Demand Processor* crawls the open data cloud to retrieve additional information to fulfill unexpected demands. This layer can be implemented using an ETL platform, which allows the implementation of needed extensions of the aDApTA.

The *Adaptive Interlinker* components are supported by a set of data stores. The first is the *Data Staging Repository* usually used to support the ETL workflow, temporarily storing the raw data extracted or received from data sources supporting the analyses, cleaning, conforming and conversion activities. In addition, the raw provenance data collected from the workflow is also stored in this repository. The second is the *Domain Triple Store* used to store the integrated domain data converted to an RDF Graph. Regarding the great volume of provenance data that can be collected, as will be shown in Chapter 5 (Application Case), another repository is used to store the provenance triples and the corresponding links to the domain data. The last data store is the *Ontology Repository* which stores the domain and provenance ontologies used by the semantic enrichment activities. These ontologies are used to evolve the conceptual schema descriptor of the data and to publish the collected provenance data, respectively. In addition, two types of supporting data are stored in this layer: the *Semantic Level Evaluation Framework*, used to identify the semantic level of the incoming data; and the *Mapping Files*, used by the *Annotate with Conceptual Description*

activity to map the vocabularies and domain ontologies to the data attributes. These two configuration files can be stored as triples or as XML machine readable file format, as exemplified in the Figure 48 and Figure 51 of the application case (Section 5.3).

The result of the process is the integrated information stored in the *Domain Triple Store* and the corresponding interlinked provenance data stored in the *Provenance Triple Store*. The other data stores are used to support the process.

Finally, the **fourth layer** comprises the available data sources with information about the specific target domain or about any other domain used to complement unexpected demands of information to support decision making. Also, the *Metadata Repository of Conceptual Reference Frameworks*, such as LOV, used by aDApTA, can be available online in the Web of Data.

All components of the *Adaptive Interlinker* layer were implemented extending the Pentaho Data Integration (PDI) platform and the triple were stored in the Open Link Virtuoso[21]. The implementation details are described in the following section.

4.3.2 ETL4LOD-Graph Framework

As explored in Section 3.5 (Supporting Tools, Frameworks and Architectures for Linking Data), there are several tools that allow the practical application of LOD principles supporting the development of integrated Web of Data applications. However, it is necessary to have features to enable the orchestration of the phases and the steps required for loading, conforming, annotating, and interlinking RDF Graphs, together with other input file formats. In addition, the provenance of the underlying interlinking approach must be collected and exposed to support data quality assessment.

The nature and variety of data sources is the primary factor to be considered when selecting the most appropriate Linked Data publishing strategy (HEATH; BIZER, 2011). The process of cleaning, transforming and integrating data from heterogeneous sources and formats to be triplified and published can naturally be orchestrated by an ETL approach (MENDONÇA et al., 2013). Such an approach uses a workflow for the linked data publishing process, inheriting the potential offered by ETL tools and techniques to perform data

[21] http://virtuoso.openlinksw.com/

preparation and integration before its triplification. Thus, implementing the *Adaptive Integration Orchestrator* of the aDApTA architecture (Figure 23), the main benefits of this approach are:

a) the systematization of the linked data publishing process;
b) the monitoring and management of the several cleaning, conforming, and integration tasks; and
c) the facilities for reusing workflows to load new data and maintain up-to-date workflow definition and data.

The main steps of the ETL workflow are:

a) extraction, when raw data is collected from sources and stored on a disk before any significant restructuring or manipulation of data takes place;
b) cleaning, when extracted data can be corrected (or rejected), and then registered, according to modifications and rules enforcement applied to the data;
c) conforming, when data is transformed according to intended purposes, from handling data duplication issues faced when merging sources to data annotation based on common vocabularies, and the conversion to a standard RDF format (triplification);
d) delivering, when data is physically structured and loaded to a destination repository in order to be further consumed; and
e) operation and management, when job execution, scheduling, exception handling, and recovery take place.

In ETL4LOD project, the Pentaho Data Integration (a.k.a. Kettle) was adopted as the workflow framework to support the LOD publishing process (CORDEIRO et al., 2011a). To implement the components of aDApTA supporting architecture, the ETL4LOD framework was extended in order to process RDF Graphs, besides rows of a relational database. This extension, called ETL4LOD-Graph, is essential to meet adaptive requirements of complex environments. As discussed in Section 3.3 (Conventional Database x Triple Stores), on relational or tabular data every row has the same schema, i.e., the same set of fields or columns. However, in dynamic environments, an information resource can be described with

different properties, even when the properties describe the same concept. Moreover, over time, new properties can be created as links are generated, improving the description of the resources.

To apply aDApTA approach, the data flow grain of the supporting ETL framework needs to be changed to a subgraph. Instead of a 3-tuple consisting of a subject, a predicate and an object, an RDF Graph is processed, representing a *Subject Item*. The RDF Graph was implemented using the interface *Model*[22] from the *Apache Jena* package[23], on which an RDF *Model* is a set of RDF *Statements*. The *Apache Jena*[24]is a free and open source Java framework for building Semantic Web and Linked Data applications.

Thus, the ETL4LOD-Graph represents another contribution of this work, where the ETL workflow engine, which natively processes relational and tabular data, was extended in order to process RDF Graphs. The major impact of this extension in aDApTA is that it allows the semantic expressivity level assessment of a subgraph composed of all statements that describe the information resource, instead of a single triple, regardless of the number of properties that composes it. This is a fundamental feature of the adaptive approach for data integration proposed in this work. Based on that, some supporting steps were implemented. Each step is described below, and their application is explored in Sections 5.3 and 5.4. Furthermore, the Appendix B gives more details about its implementation.

[22] http://jena.apache.org/documentation/javadoc/jena/com/hp/hpl/jena/rdf/model/Model.html

[23] https://jena.apache.org/documentation/javadoc/jena/jena/package-summary.html

[24] https://jena.apache.org/

RDF Graph SPARQL Query: the first step of an ETL workflow is the extraction of data from a source. When the source is a database, a query step is used. In our case, the source can be a SPARQL endpoint. In the first version of ETL4LOD, the step *SPARQL Query* was developed to retrieve triples and process the *SELECT* command results, which has a columnar format. In order to retrieve RDF Graphs from a SPARQL endpoint, this step was improved. It now runs a SPARQL query against an endpoint and retrieves a set of RDF Graphs composed of triples. Thus, the SPARQL commands can only be *DESCRIBE* or *CONSTRUCT*. The input parameters are: SPARQL endpoint URL, Named Graph and SPARQL Query. The output is an object composed of a set of RDF Graphs. Figure 24 presents a screen shot of this step configuration interface where the *Query* tab is selected and shows the SPARQL query input and validation.

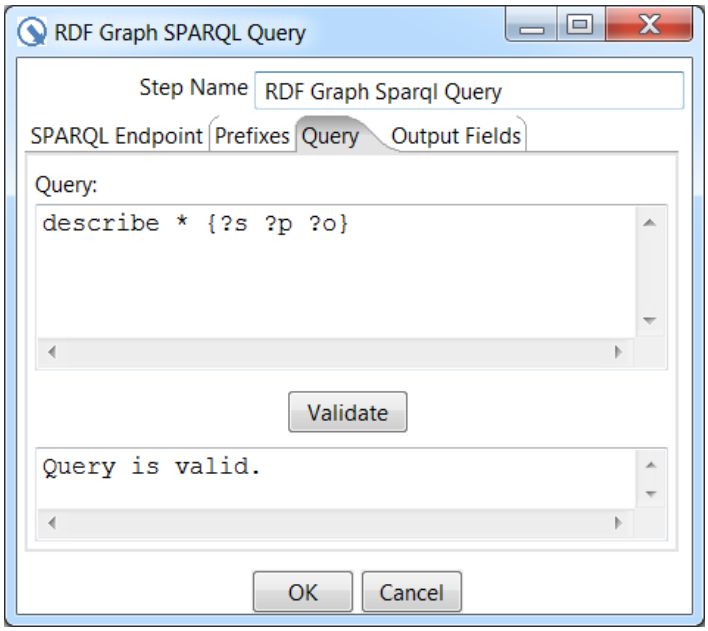


Figure 24: Configuration interface of the *RDF Graph SPARQL Query* step

RDF Graph Semantic Level Marker: in aDApTA, the semantic level of an RDF Graph is evaluated and maintained on the database to support the definition of a corresponding interlinking tool. To do that, this step reads an RDF Graph, evaluates its semantic expressivity level and creates a new triple, stamping its level. It is performed as described in Algorithm 1 (see Section 4.1.1). The input is the *Graph Field* representing the RDF Graphs retrieved by the *RDF Graph SPARQL Query* step. The outputs are the subject, predicated and object fields which compose a *stamp triple* describing the semantic level of the RDF Graph. The configuration parameters are the location of the XML *Linked Open Vocabulary* file and the location of the XML *Semantic Level Evaluation Framework* file. The rules defined in the framework are performed in Java code with the *eval* method of the *Interpreter Class* of the *BeanShell* package[25]. The Figure 25 presents the screen shot of the configuration interface which allows the filling of the parameters.

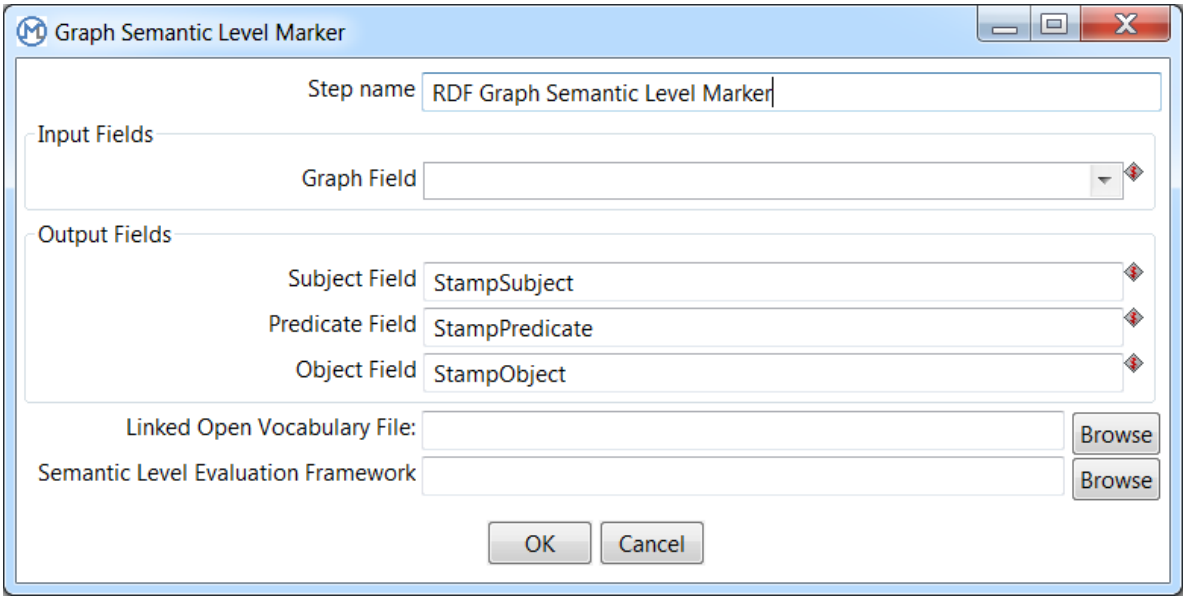


Figure 25: Configuration interface of the *RDF Graph Semantic Level Marker* step

[25] http://www.beanshell.org/javadoc/bsh/Interpreter.html

RDF Graph Triplifier: during the transformation process, there are cases where each triple, which composes the RDF Graph, must be processed individually. Thus, this step was developed to read an RDF Graph and return a set of rows with three fields containing the subject, predicate and object of each triple of the RDF Graph. In short, this step is used to change the grain of the data processed by the workflow from RDF Graph to triples. Figure 26 presents the screen shot of the configuration interface of this step.

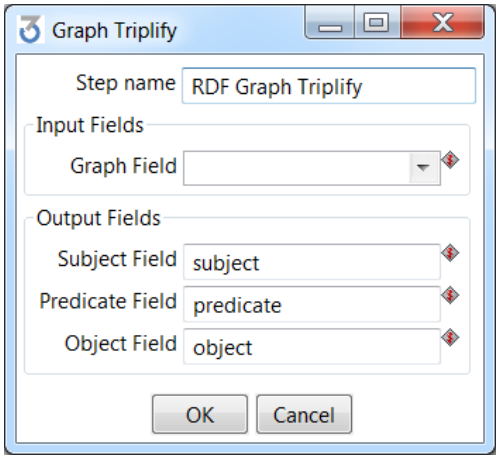


Figure 26: Configuration interface of the *RDF Graph Triplifier* step

Triple Annotator: in order to improve the semantic expressivity level of an RDF Graph, this step annotates the triples with a conceptual description model. As described in Algorithm 2 (Section 4.1.2), it automatically changes the property or the object of a triple with terms from vocabularies and ontologies according to a mapping FROM-TO XML file, built manually. The input parameters are triples and the location of the XML mapping file. The output is the annotated triple. Figure 27 shows the screen shot of the configuration interface where the parameters are filled together with the path of the *Mapping file*.

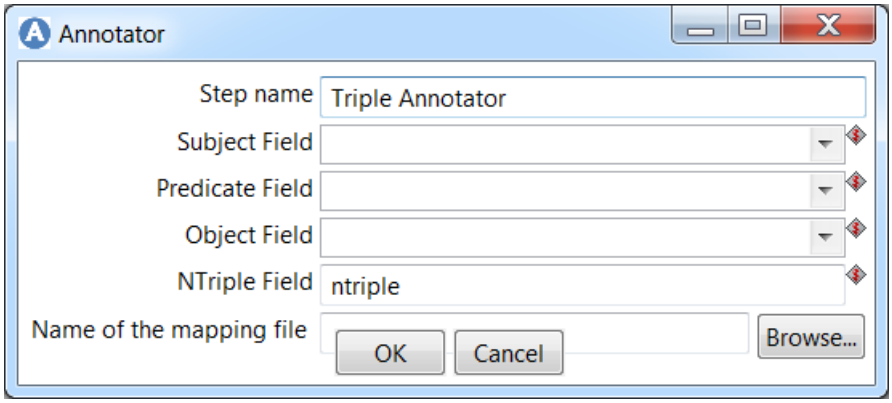


Figure 27: Configuration interface of the *Triple Annotator* step

The screen shots of Figure 28 and Figure 29 illustrate the Pentaho Data Integration (PDI) development environment, which is composed of a pallet listing the available steps, located on the left side of the screen, and a workspace, located on the right side, where the workflow is developed with a set of configured steps. In Figure 28, the new steps of ETL4LOD-Graph are illustrated on the pallet and on the workspace. Likewise, Figure 29 illustrates the improved steps of ETL4LOD.

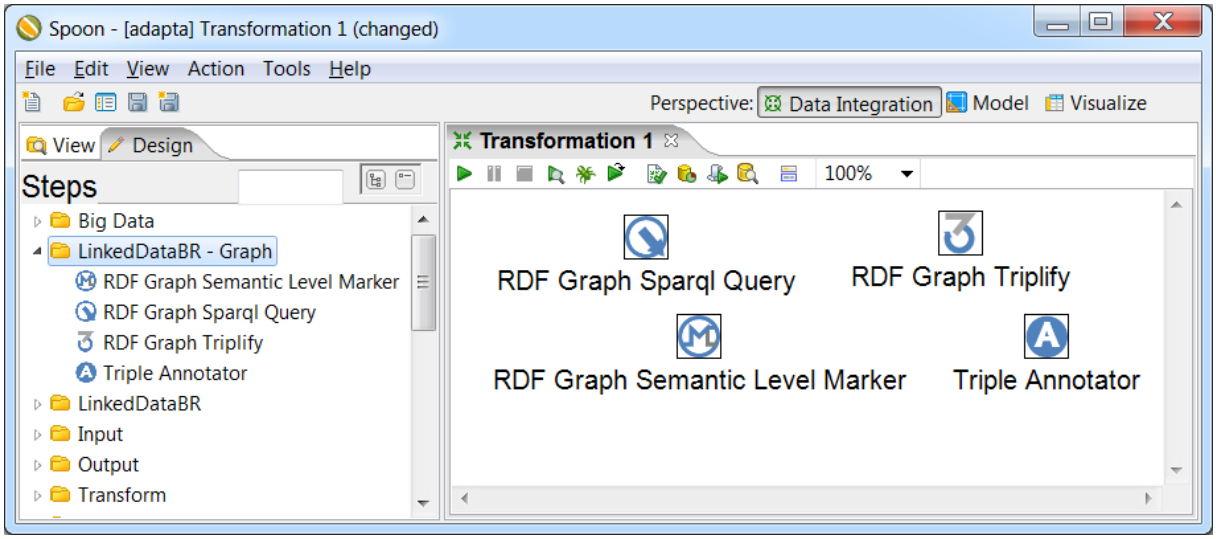


Figure 28: PDI pallet of ETL4LOD-Graph steps

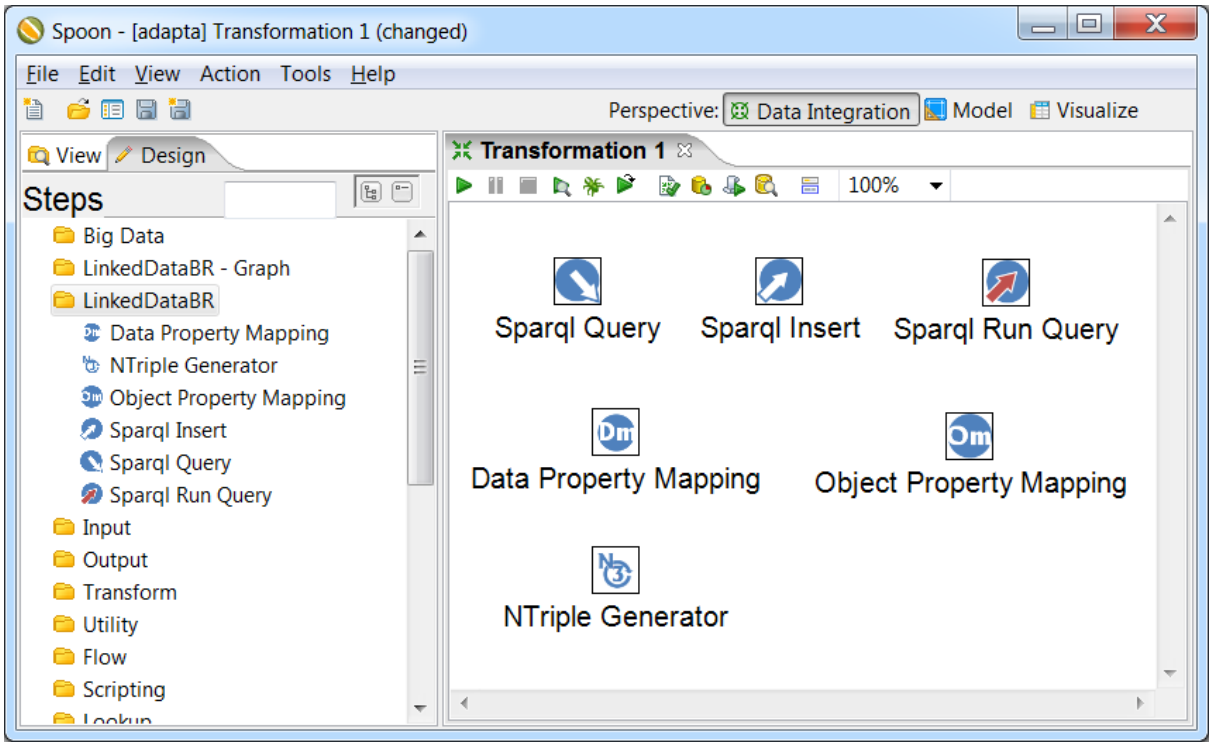


Figure 29: PDI pallet of improved ETL4LOD steps

The improvement on ETL4LOD was the development of the **SPARQL Run Query** (Figure 30) which enables the specification of a text query to perform updates in the dataset, such as *UPDATE DATA, DELETE DATA* and *DROP GRAPH*. Another minor improvement was the design and standardization of all steps' icons.

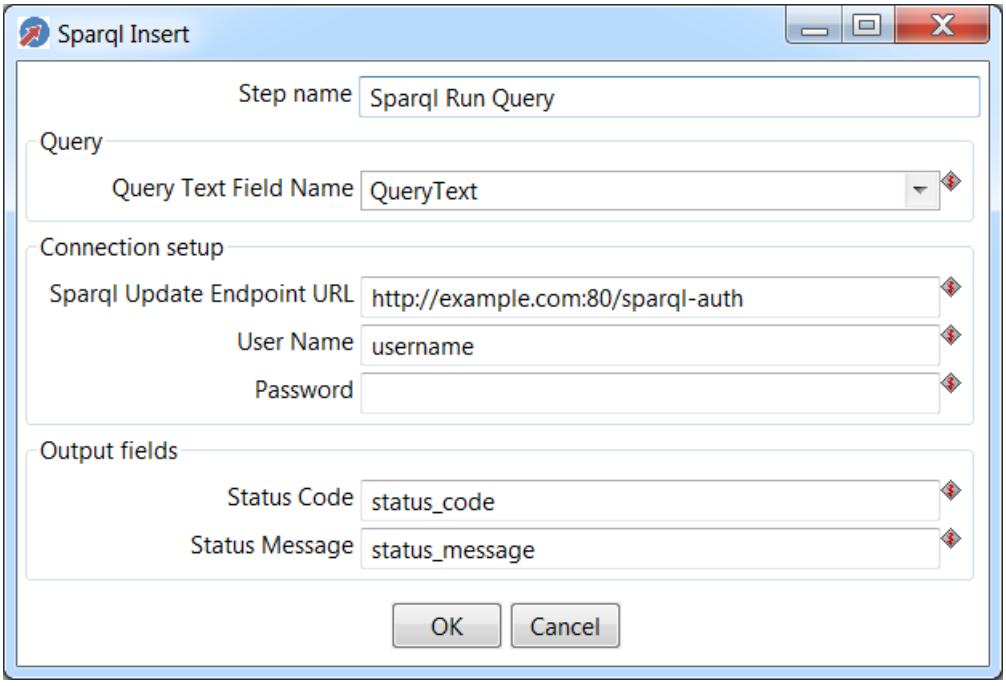


Figure 30: Configuration interface of the *SPARQL Run Query* step

In order to enable the provenance collection of the new steps described above, the Provenance Agent of the ETL4LinkedProv[26] (see Section 3.5) was extended.

Provenance Collector Agent: this step captures the prospective and retrospective provenance data, and makes the interlinking to the transformed data from the domain, during the execution of the transforming process. In fact, this is a special step called job in PDI. The window configuration of the Provenance Collector Agent is displayed in Figure 31. The new steps are marked on the step type list informing the agent to collect their fine-grained provenance. The PDI pallet of extended ETL4LinkedProv Agent is presented in Figure 32.

[26] https://github.com/rogersmendonca/provenance_collector

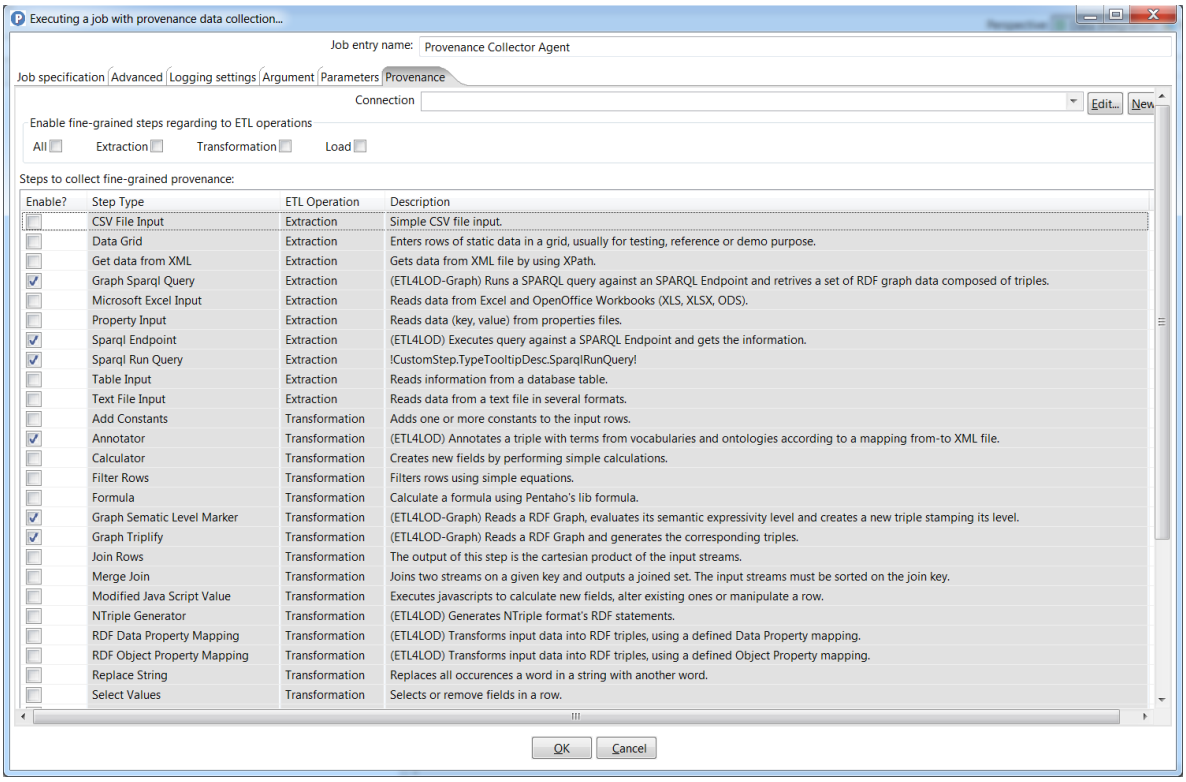


Figure 31: Configuration interface of the *Provenance Collector Agent* with ETL4LOD-Graph steps

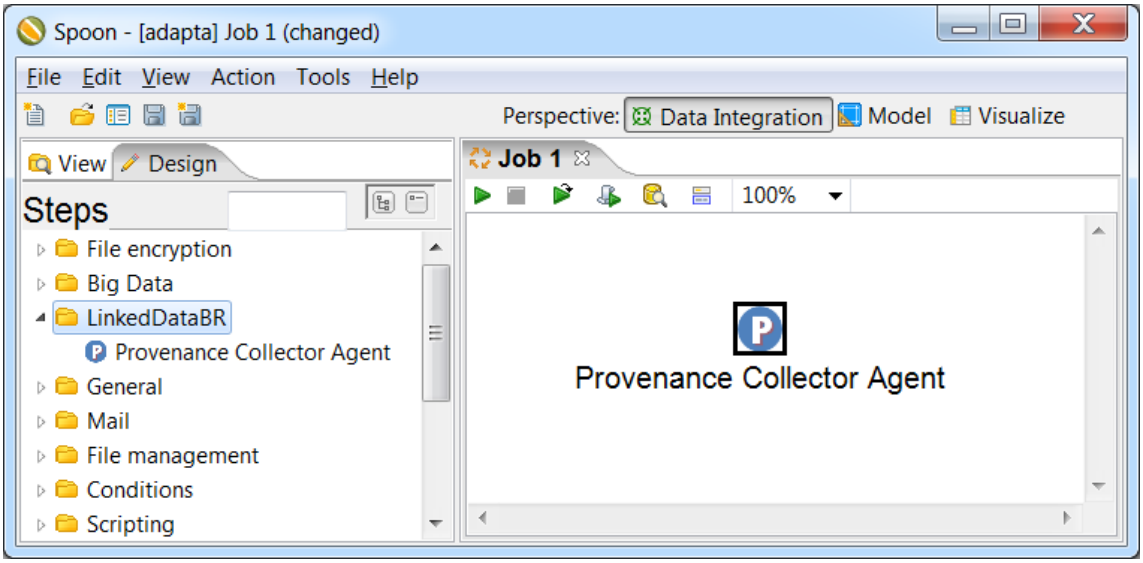


Figure 32: PDI pallet of extended ETL4LinkedProv agent

The source code of ETL4LOD and the presented extensions are available on gitHub[27]. They are ready to be applied in any domain within a user-friendly interface, as presented in the application case (Chapter 5).

[27] https://github.com/Kelli

4.4 Preliminary Conclusions

This chapter presented the main contribution of this thesis, an approach for adaptive integration of information to support decision making in a complex environment, called aDApTA. The strategy is to use the flexibility of graph data models to solve structure conflicts by converting the data to an RDF Graph and to apply the LOD principles for information integration. Based on that, it is possible to identify information resources described with different semantic expressivity levels that can prevent or compromise the information integration results. The importance of identifying and solving this type of conflict is growing because of the different ways of publishing data on the Web.

The aDApTA activities, strategy and architecture were detailed showing how the identification of semantic expressivity level heterogeneity enables system adaptation by switching the integration approach to the most appropriate one. In order to illustrate the addressed problems and the proposed solution, a real scenario was used as an application case, as described in the following chapter.

5 Application Case

In order to evaluate the proposed approach, the emergency management domain was chosen because it has the main characteristics of a decision making process under a dynamic environment. Emergencies are typical complex environments, where several agents, such as persons, organizations, equipment and systems must constantly adapt to an unpredictable evolving situation.

Emergency management can be presented as a cyclic process composed of three phases: (a) pre-disaster, which starts before an emergency event and aims to define policies and actions to reduce the vulnerability of a population or minimize the adverse effects of future disasters; it involves actions of mitigation (e.g., building and zoning codes; vulnerability analysis, etc.) and preparedness (e.g., elaboration of emergency plans; emergency exercises and training; warning systems deployment, etc.); (b) response, which starts when a dangerous situation occurs requiring immediate action. During this phase, response teams perform actions to reduce the negative effects caused by the event; (c) post-disaster, which starts when the emergency is controlled, and aims to repair, rebuild, and restore what was lost during the disaster; it can take weeks, months or even years to be concluded (KHAN et al., 2008; HADDOW et al., 2011). Moreover, the actions can be strategic, tactical, or operational. The jurisdiction of the activities may include different hierarchical levels, such as federal, state, local or the equivalent (LINDELL et al., 2006).

Information systems support the activities of emergency management phases and levels facing several challenges, especially those that support the response phase at the strategic level. A large volume of information at different levels of detail is collected and evaluated to define further actions. Therefore, the information base architecture and management are critical to the success of an emergency response. In the response phase of emergency management, information acts as a crucial tool and has unique characteristics that require specific handling for an effective use of its content. The information available to decision makers can be derived from internal or external knowledge bases, represent current or previous knowledge, and its nature can be static or dynamic (DINIZ et al., 2005).

During this phase, the supply and demand of information changes continuously according to contextual changing. Both the supply and demand for information can only be partially predictable, requiring managers to constantly adapt their resources and data flows to keep decision makers informed (BHAROSA; JANSSEN, 2010; HELLINGRATH; WIDERA, 2011).

Based on interviews with domain experts and simulation exercises with emergency response teams, the main challenges of information management are (SIAU; TIAN, 2004; BHAROSA; JANSSEN, 2010; BARR et al., 2010; HRISTIDIS et al., 2010; BARR et al., 2011): (i) to gather, integrate, and reconfigure the internal and external resources; (ii) to disseminate reliable and timely information; (iii) to know which information exists and who controls it; (iv) to obtain relevant information or to add value to it. Regarding the sources of information, the authors suggest that the amount of alternative sources should be maximized.

Among several challenges of the emergency domain, the application case focuses on: (i) unpredictability of data sources and demands; (ii) integration of information with structural, semantic and temporal heterogeneity; and (iii) the consequent uncertainty about the quality of data in a dynamic and heterogeneous environment. One emergency case was chosen to apply aDApTA as described in the following section.

5.1 Application Case Scenario

In November 2013, the category 5 Typhoon Haiyan made a direct hit on the Philippines. Many cities experienced widespread destruction, reaching 90% of housing destroyed in some areas. Roads were blocked, and airports and seaports became impaired. Water supply and power were cut, food stocks and other goods were destroyed, many health facilities crashed, and medical supplies were quickly being exhausted. To aid the victims of affected areas, many organizations worldwide were mobilized. The government in the Philippines identified food, water, sanitation and hygiene, shelter, medicine, debris clearing, and logistics as immediate priorities. It also requested the international community's support in establishing logistics hubs to allow the sustainable delivery of aid (UNDP, 2013; UNOCHA, 2013). Furthermore, to optimize the use of available transportation facilities among more than 7000 islands, a complementary information demand emerged: the location and distances between the warehouses and the beneficiaries. To meet this

demand, the latitude and longitude coordinates could be obtained from open data sources, such as DBPedia[28] (LEHMANN et al., 2013) and GeoNames[29].

Figure 33 depicts the scenario of this case where adaptive procedures are demanded in response to the dynamic behavior of the emergency. Additionally, the figure highlights the aDApTA context.

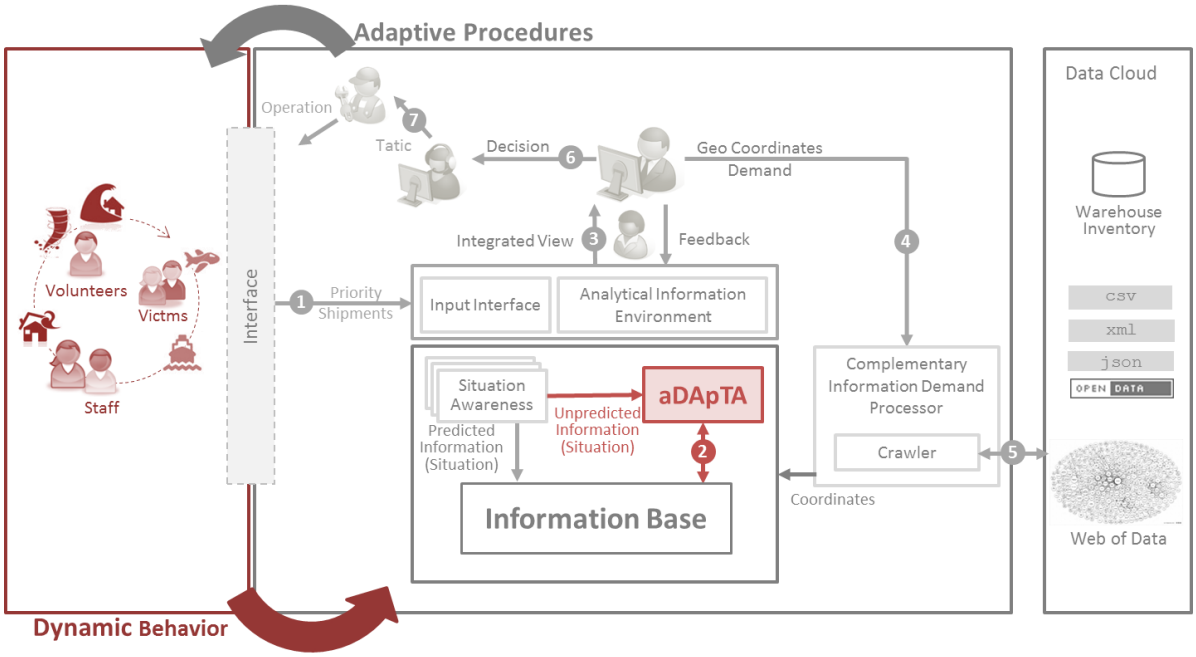


Figure 33: Overview of the application case scenario

The humanitarian logistics process in disaster relief supply chains can be divided into four phases (WIDERA et al., 2013):

a) *assessment*, the process of gathering, analyzing and disseminating logistics related data and information in relation to the disaster impact;

b) *procurement*, the process of identifying and obtaining goods and services from potential suppliers;

c) *warehousing*, the process for storing and handling goods and materials, being an interface between suppliers and beneficiaries;

d) *transport*, comprises the activities involved in physically moving supplies in a reliable and safe manner, on time, cost effectively and efficiently to its destination.

[28] http://dbpedia.org/

[29] http://geosparql.org/

These phases form a cycle where the current situation is observed and reported by the team facing the event (1). When the report contains information with characteristics that were not predicted, aDApTA is applied to integrate the information (2). Then, the decision maker analyzes the unified view of the situation (3). If the information is not enough, the decision maker can demand complementary information (4). This is done by aDApTA, crawling the Web of Data Cloud (5) and integrating the information (2). Based on this information, the decision made at the strategic level (6) is passed on to the tactics team and performed by the operational team (7), affecting the situation. Thus, new analyses of the situation are required and new decisions will need to be made, restarting the cycle. The application of the adaptation strategy for information integration in this scenario is described in detail in the next section. Furthermore, the implementation details, links to the files and an user guide is provided in Appendix B.

5.2 Data Sources Profile

In order to provide integrated and reliable information about the ongoing situation, the first activity of aDApTA is performed, the *Analyze Structure of Input Information* (see section 0). The analysis starts with the generation of a data profile.

Initially, two data sources were used. The first one is the relational database of Sahana Eden Open Source Platform[30] that supports the management of data about the shipments of goods from storehouses to beneficiaries. It is an initiative of the Philippines Government and Sahana Software Foundation[31]. Many worldwide entities, governments and volunteers feed the system with information through Web forms. This information is available in xls files on system web site (http://eden.dswd.gov.ph/eden/).

The data model of Sahana Eden Platform deployed to support the Philippines disaster logistic is presented in Figure 34. It was designed using reverse engineering and is composed of five tables representing five entities. The warehouse and facilities represent locations where the relief goods are stored. They are managed by organizations that keep the

[30] http://eden.sahanafoundation.org/

[31] http://sahanafoundation.org/

storehouse. The items of first aid are grouped in kits to be distributed. The information about the kits received and sent to warehouses are stored in the shipment table.

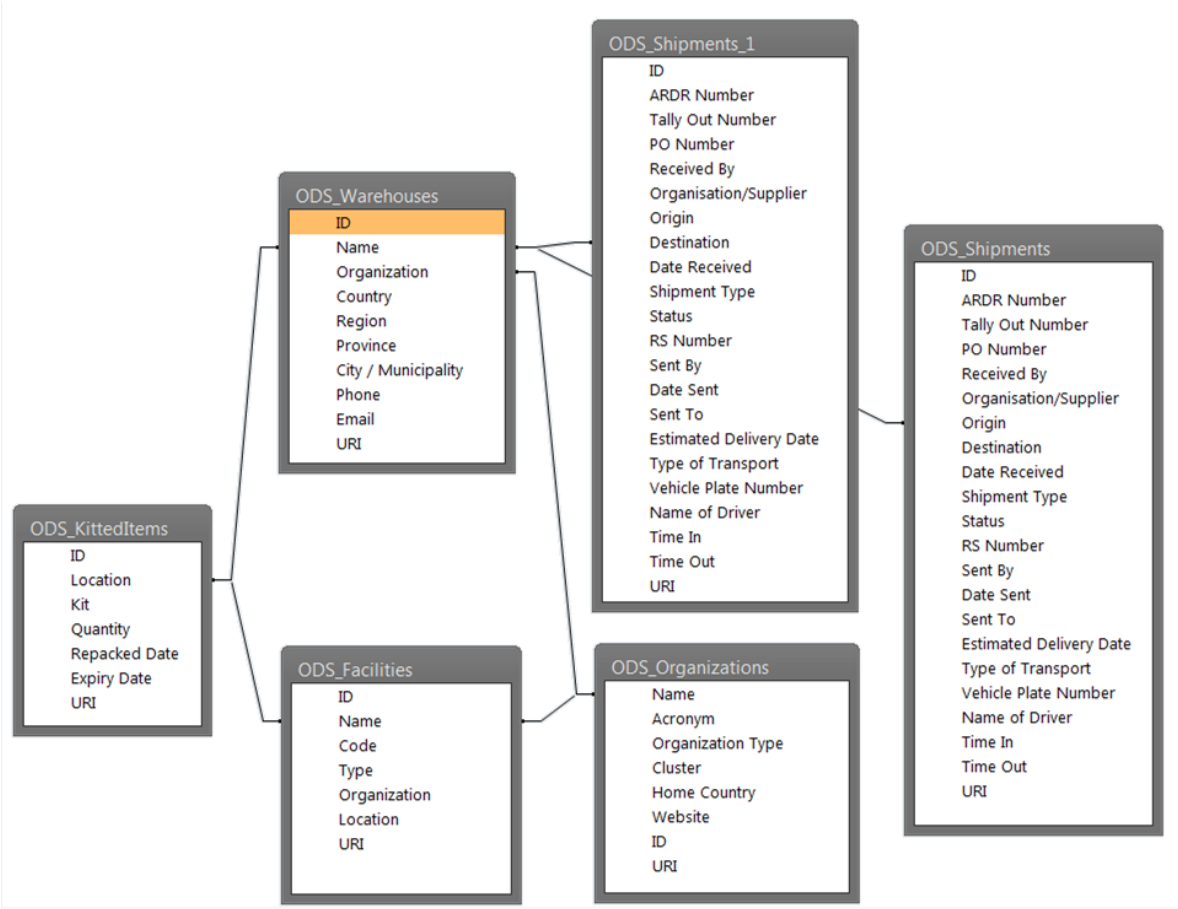


Figure 34: Sahana Eden data model

The second data source is a set of spreadsheets managed by the United Nations (UN). The information about the donations of worldwide senders are collected by UN staff and made available on a Web site (http://logik.unocha.org/SitePages/map.aspx). In addition, the United Nations Office maintains a spreadsheet file about the population and displaced persons in each municipality for the Coordination of Humanitarian Affairs (UNOCHA) and the Department of Social Welfare of Philippines (DSWD), to support the decision making about priority regions (UNOCHA, 2014). The data model of OCHA tabular files is illustrated on Figure 35. It was also designed using reverse engineering and is composed of four tables. The relied items and contributions store information about the donations. The transport table records information about the vehicles used to send the relief goods, and the priority regions affected by the typhoon are stored in the municipalities table.

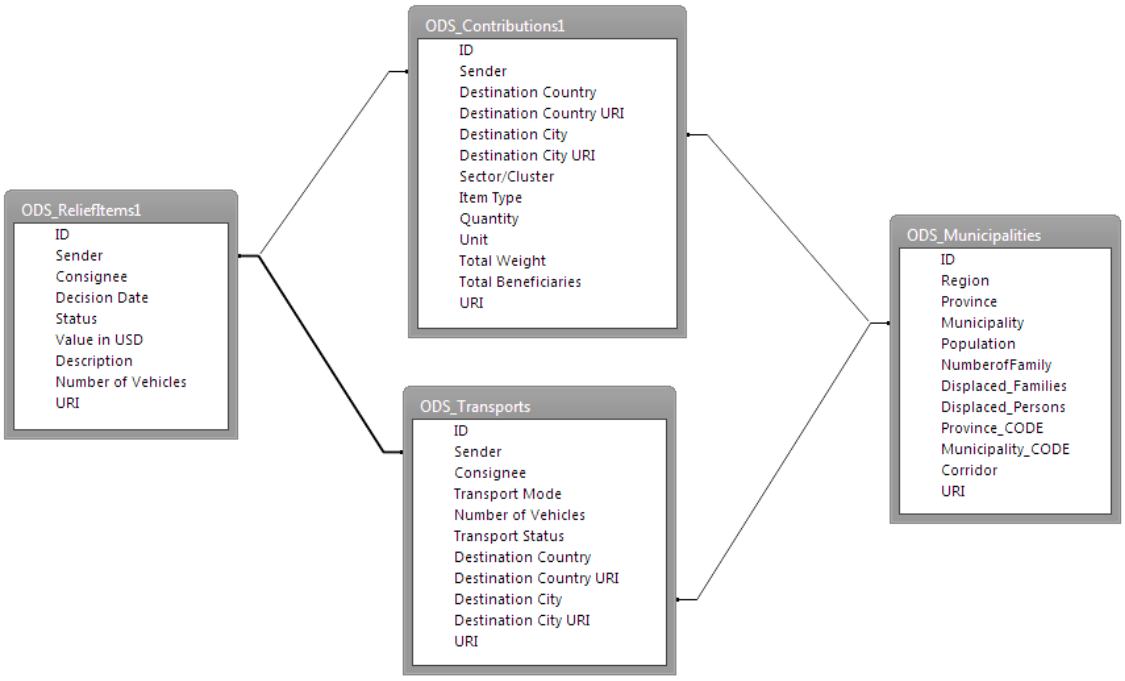


Figure 35: OCHA data model

After downloading the source data files, some basic profiling tasks were performed, such as the count of tables, records, empty and comment fields, plus the period of the time dimension. Additionally, a semantic profiling was done counting distinct terms and correlating them. Figure 36 shows the statistics results.

Source	Table	Record	Field	Cell	Filled	Empty	Comment
Sahana Eden	5	4.036	45	114.351	86.616	27.735	1.964
OCHA	4	627	23	4.070	3.627	443	-

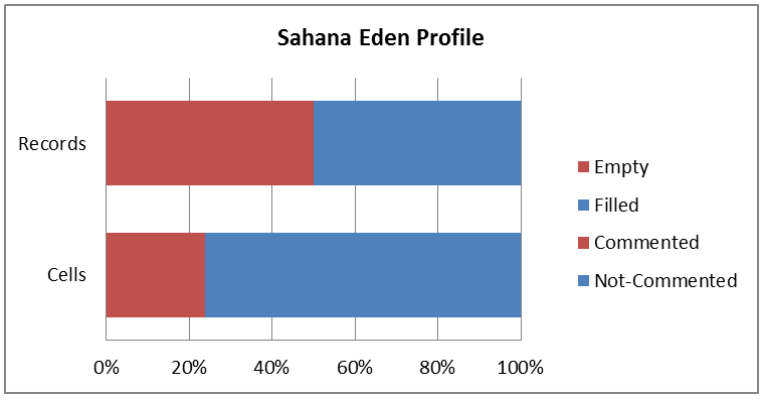


Figure 36: Data source profiling

Analyzing the Sahana Eden raw data, almost 30% of the cells are empty and 50% of the records used the comment field. These scores indicate that a great portion of the database schema is not suitable for the situation. On the other hand, the OCHA/DSWD spreadsheet is completely filled; however, there were many special characters that demanded some extra cleaning steps. Besides the data count, the time dimension was analyzed. It reveals that in the first days of the event, only a few received shipments were made, and later, it increased. Furthermore, it indicates that the assessment of the situation takes a long time.

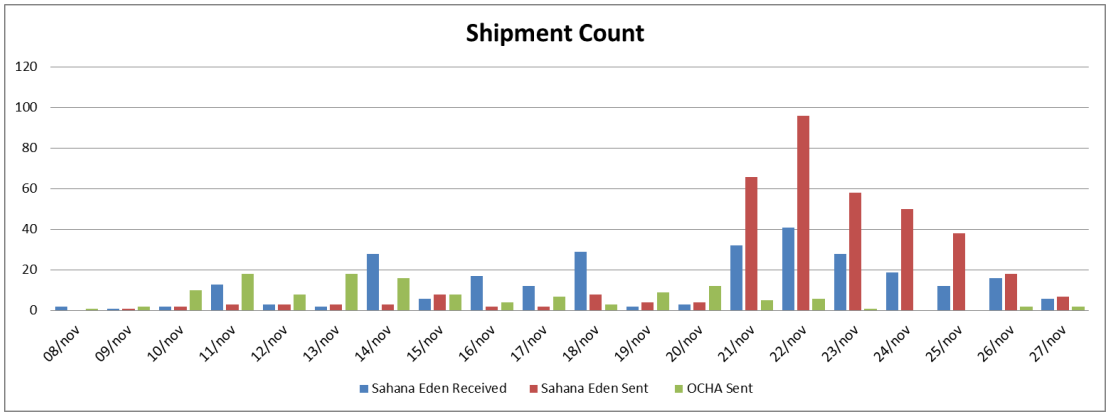
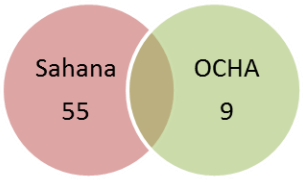


Figure 37: Time dimension analyses of data source

From a semantic perspective, there are only six concepts used by OCHA and Sahana Eden as well (see Figure 38a). Only three concepts use the same terms to express them (see Figure 38b). This indicates the great semantic distance between the datasets increasing the difficulties of integration tasks. To minimize this issue, a domain ontology can be used to map the concepts between the schemas. The development of the domain ontology and its use with the data sources are described in the following sections.



a

Sahana Eden	OCHA
Cluster	Sector/Cluster
Date Sent	Decision Date
Quantity	Quantity
Sent By	Sender
Status	Status
Type of Transport	Transport Mode

b

Figure 38: Terms and concepts count of data sources

5.2.1 Domain Ontology Development

To semantically enrich the data of the source files, a domain ontology was used. The *Humanitarian Logistic Domain Ontology* was developed using a hybrid approach. The data sources concepts were analyzed and used to compose the first version of the ontology. Then a core ontology, proposed by Giannotti (2011), was used to complement the missing concepts. Moreover, the following standard vocabularies and ontologies were reused to describe time and space concepts:

- Time (http://www.w3.org/TR/owl-time)
- Geo (http://www.w3.org/2003/01/geo/wgs84_pos)
- GeoNames (http://www.geonames.org/ontology)
- Places (http://purl.org/ontology/places)

The final ontology is composed of 47 domain concepts and 19 concepts about time and location. The Figure 39 provides a general overview of the ontology diagram. The yellow rectangles represent the location concepts, the blue rectangles represent the time concepts, and the red rectangles represent the main domain concepts. The detailed diagram and the OWL code of the developed ontology can be found in Appendix A.

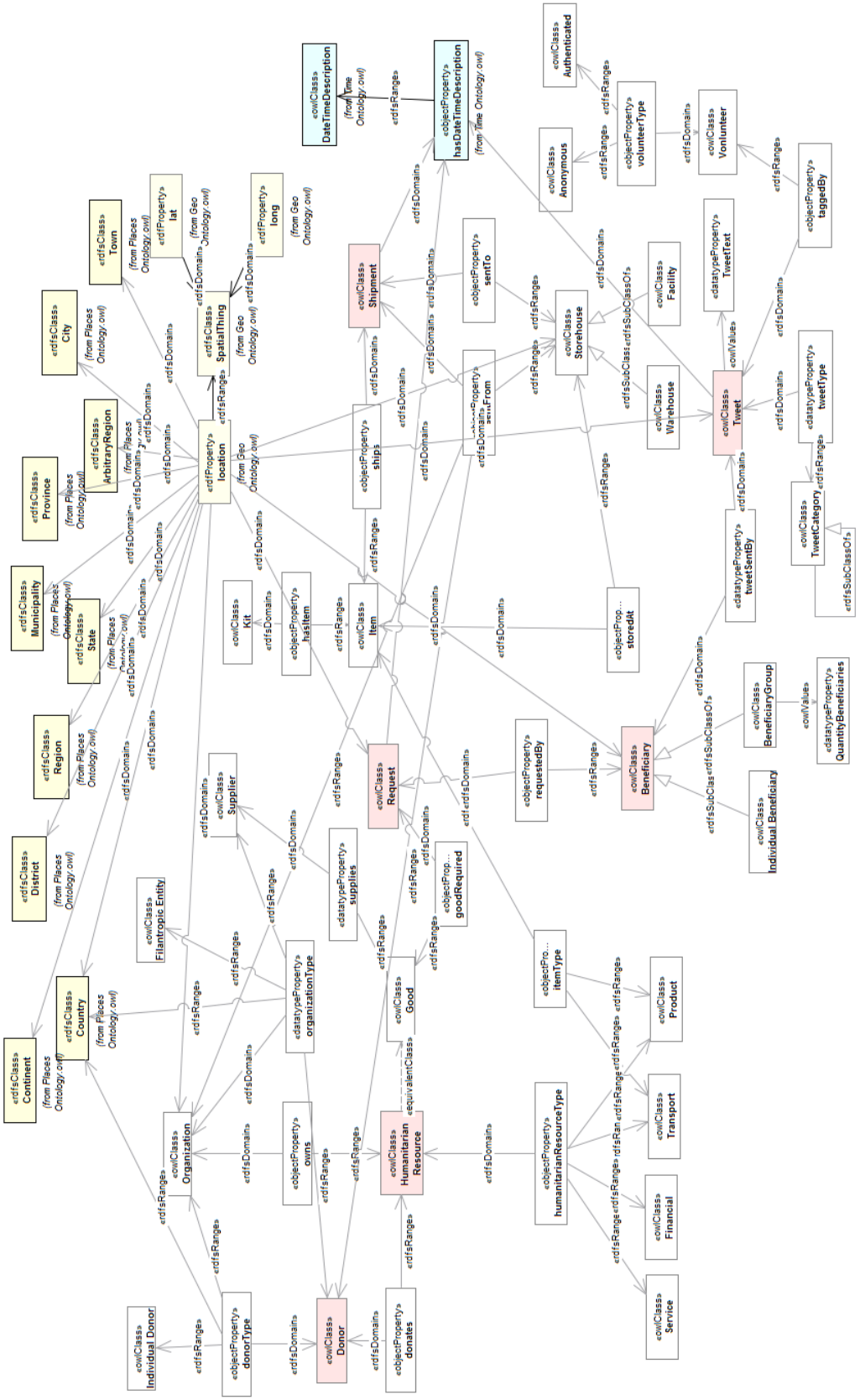


Figure 39: OWL diagram of humanitarian logistic domain ontology

5.2.2 Data Source Scope

A small set of the data sources was used to exemplify the application case. It is highlighted in a screen shot of the Sahana Eden Web form displayed in Figure 40, and in the spreadsheet presented in Figure 41, with information about the population and displaced people in each municipality in the Philippines.

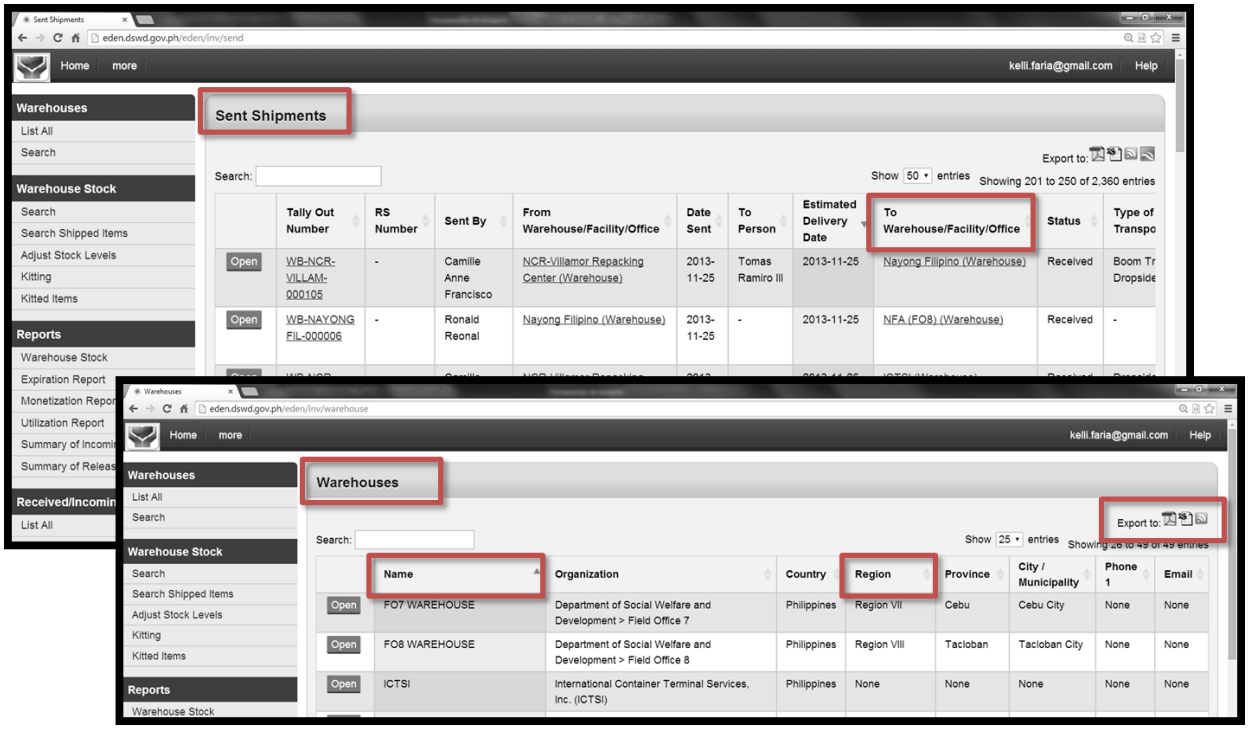
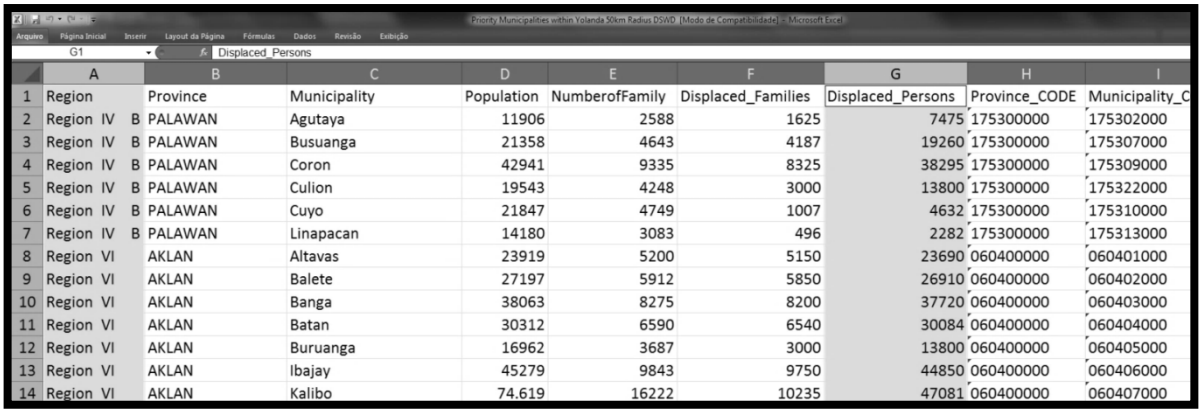


Figure 40: Sahana Eden data source (http://eden.dswd.gov.ph/eden/)



	A	B	C	D	E	F	G	H	I
1	Region	Province	Municipality	Population	NumberofFamily	Displaced_Families	Displaced_Persons	Province_CODE	Municipality_C
2	Region IV B	PALAWAN	Agutaya	11906	2588	1625	7475	175300000	175302000
3	Region IV B	PALAWAN	Busuanga	21358	4643	4187	19260	175300000	175307000
4	Region IV B	PALAWAN	Coron	42941	9335	8325	38295	175300000	175309000
5	Region IV B	PALAWAN	Culion	19543	4248	3000	13800	175300000	175322000
6	Region IV B	PALAWAN	Cuyo	21847	4749	1007	4632	175300000	175310000
7	Region IV B	PALAWAN	Linapacan	14180	3083	496	2282	175300000	175313000
8	Region VI	AKLAN	Altavas	23919	5200	5150	23690	060400000	060401000
9	Region VI	AKLAN	Balete	27197	5912	5850	26910	060400000	060402000
10	Region VI	AKLAN	Banga	38063	8275	8200	37720	060400000	060403000
11	Region VI	AKLAN	Batan	30312	6590	6540	30084	060400000	060404000
12	Region VI	AKLAN	Buruanga	16962	3687	3000	13800	060400000	060405000
13	Region VI	AKLAN	Ibajay	45279	9843	9750	44850	060400000	060406000
14	Region VI	AKLAN	Kalibo	74.619	16222	10235	47081	060400000	060407000

Figure 41: United Nations (UNOCHA) and the Department of Social Welfare of Philippines (DSWD) data source

Thus, in this scenario, the data sources used in the application case have the characteristics described in Table 5.

Table 5: Dataset scope of the application case

Source	Data Items	Format	Application Case Scope
Sahana Eden	45 fields 4,000 records	Relational	Shipment and Warehouse
OCHA/DSWD	4,000 cells	Spread Sheet	Priority
Humanitarian Logistic Ontology	47 domain concepts 19 time and location concepts 12 location instances	OWL	Shipment, Beneficiary, Location

5.3 ETL Workflow of the Adaptive Interlinking

After the analysis of the data sources profile and the domain ontology modeling, the ETL workflow process was developed via the steps of the ETL4LOD-Graph framework (section 4.3.2). Each activity of aDApTA was implemented through a transformation package. All packages were put together in a job package representing the *Adaptive Integration Orchestrator* (Section 4.3.1), as shown in Figure 42.

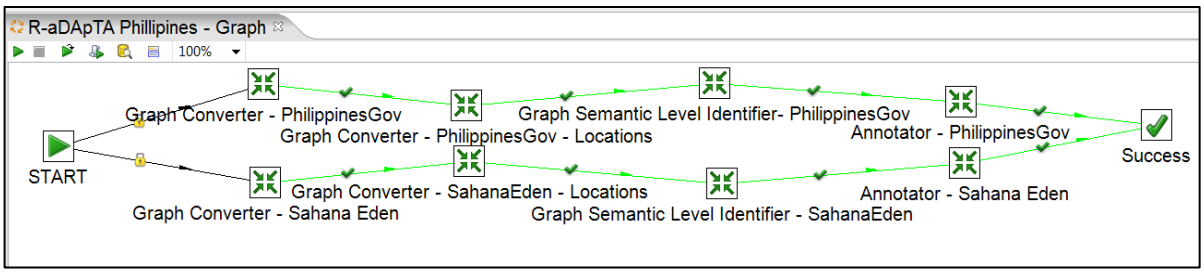


Figure 42: Job with the set of aDApTA transformations package applied to the case

The workflow starts with the *Convert to RDF Graph* activity that was implemented through the transformation package displayed in Figure 43 and Figure 44. The xls files were cleaned to remove special characters, titles, blank lines, and spaces; an URI was defined to compose the subject; the properties were also defined; and, finally, the triple was generated and stored in a triple store. To implement these last tasks, the ETL4LOD steps were used together with the native PDI steps. A sample of the *Graph Converter Package* result is illustrated in Figure 45.

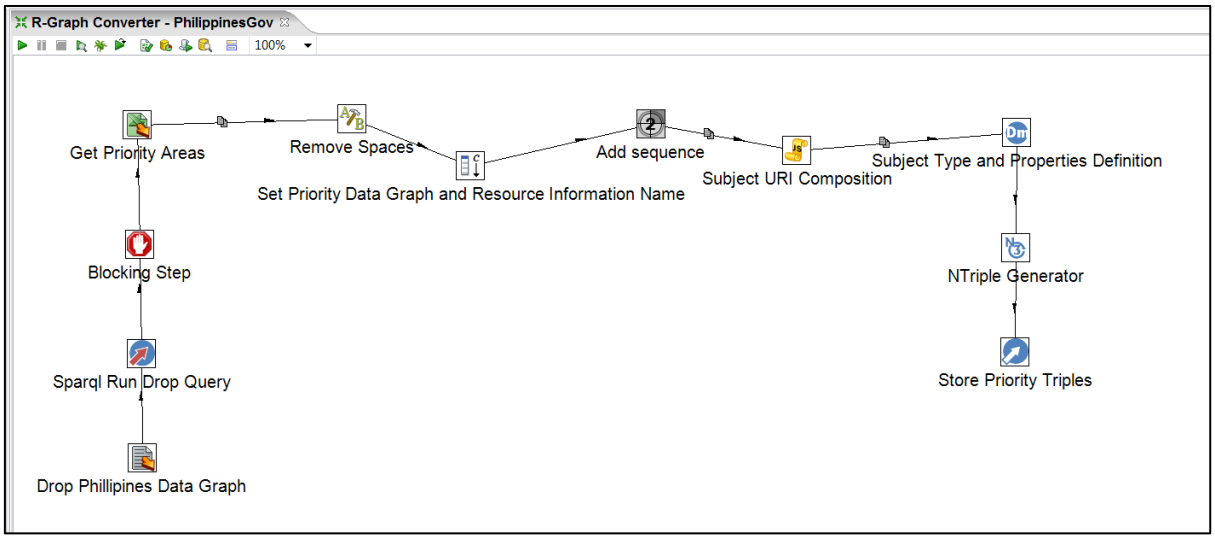


Figure 43: *RDF Graph Converter* transformation package of Philippines priority locations

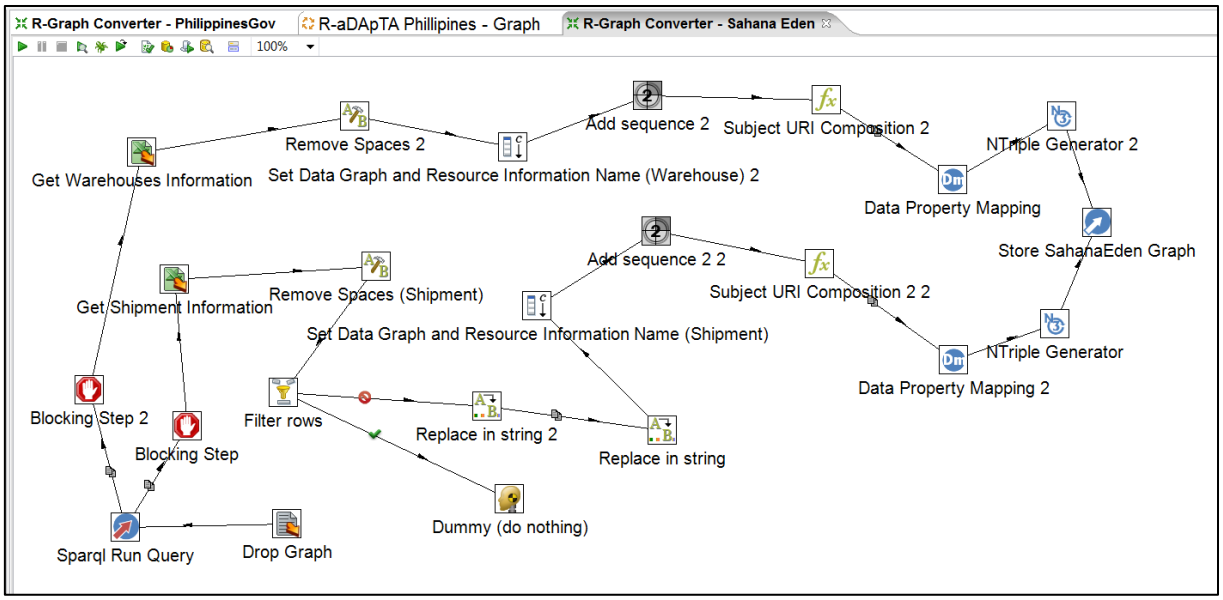


Figure 44: *RDF Graph Converter* transformation package of Sahan Eden data

```
<shipment254><towarehousefacilityoffice>"FO7 WAREHOUSE"
<warehouse26><type><Warehouse>
<warehouse26><name>"FO7 WAREHOUSE"
<warehouse26><region>"Region VII"
<priority101><region>"REGION VII"
<priority101><displaced_persons>74785
```

Figure 45: Subset of triplified data generated by the *RDF Graph Converter* package.

After the data sources conversion to RDF Graph, one of the main activity of aDApTA is the identification of information resources semantic level. To implement this activities, the step *RDF Graph Semantic Level Marker* was used. The screen shot displayed in Figure 46 shows the implementation and the input parameters of this step. A sample of the package execution result is illustrated in Figure 47.

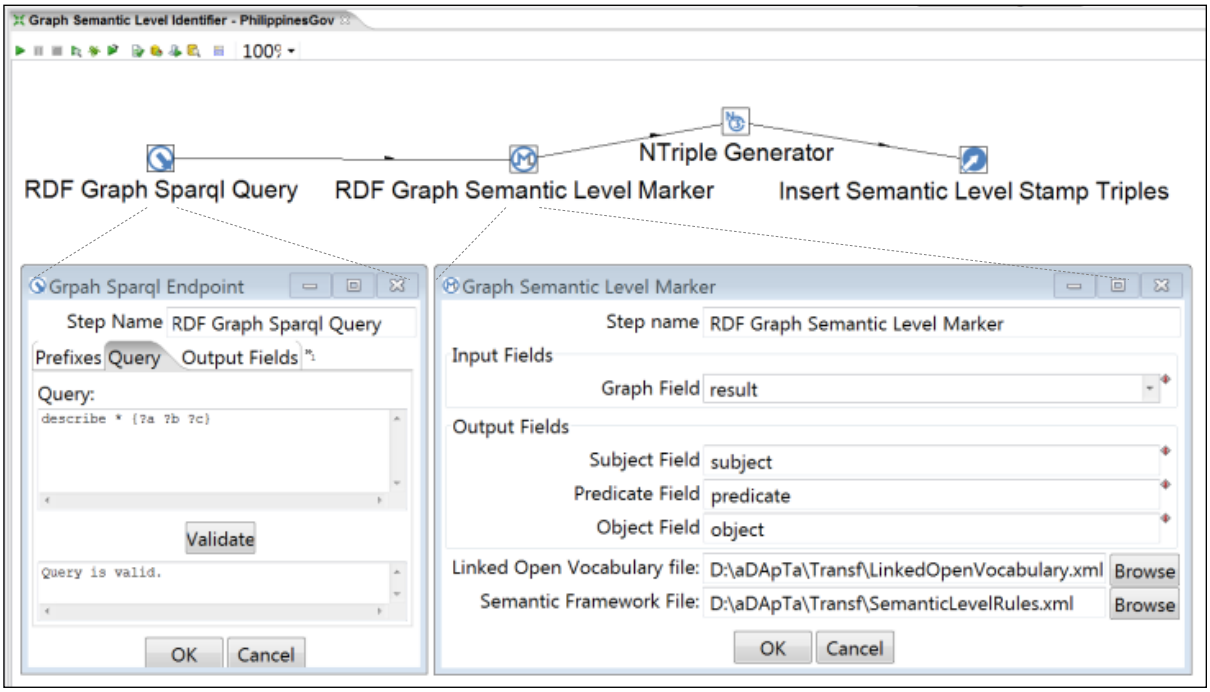


Figure 46: *RDF Graph Semantic Level Identifier* package of the Philippines priority locations

```
<shipment254><sstamp:hasSemanticLevel><sstamp:low>
<warehouse26><sstamp:hasSemanticLevel><sstamp:medium>
<priority101><sstamp:hasSemanticLevel><sstamp:low>
```

Figure 47: Sample of *stamp triples* generated by the *RDF Graph Semantic Level Marker*

The *RDF Graph Semantic Level Marker* step has two input files. The first one is the *Metadata Repository of Conceptual Reference Frameworks*. In the case, the LOV repository was used (see Section 4.1.1, Identify Semantic Level). The second input file is the *Semantic Level Evaluation Framework.* It is presented in Figure 48 and has 3 basic levels.

```
<?xml version="1.0" encoding="UTF-8"?>
<SemanticLevelFramework>
    <Frame id="1">
      <Rule>s.getLiteral() != null</Rule>
      <LevelValue>1</LevelValue>
      <LevelDescription>sstamp:low</LevelDescription>
   </Frame>
   <Frame id="2">
      <Rule>isVocabulary</Rule>
      <LevelValue>2</LevelValue>
      <LevelDescription>sstamp:medium</LevelDescription>
   </Frame>
      <Frame id="3">
      <Rule>isOntology</Rule>
      <LevelValue>3</LevelValue>
      <LevelDescription>sstamp:high</LevelDescription>
   </Frame>
</SemanticLevelFramework>
```

Figure 48: *Semantic Level Evaluation Framework* applied to the case

Having the information resources marked with different semantics expressivity (see Figure 47), the following activity is solving the semantic conflict by annotating the triples with a conceptual description, in the case, a Domain Ontology. The *Humanitarian Logistic Ontology* used in the application case is illustrated in Figure 49. The concepts of the example are highlighted. The *Beneficiary Group* class and *Beneficiary Quantity* property represent the concepts of the OCHA/DSWD displaced person priorities. The *Shipment* class and *sentTo* property represent the Sahana Eden shipments. The *Storehouse*, which can be a *Warehouse* or a *Facility*, represents the destination of the shipments. The property *Location* of the *Region* class is used to represent both shipments destination and priority location, enabling the composition of an integrated view of the situation.

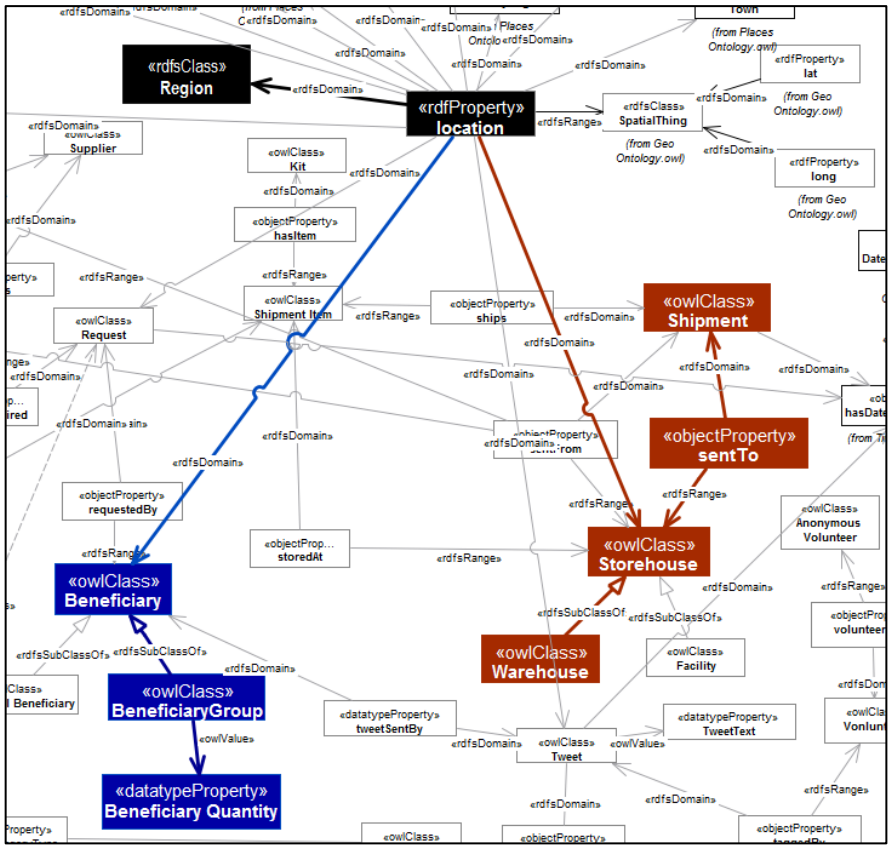


Figure 49: Diagram of OWL domain ontology of the scenario case

To perform the activity *Annotate with Conceptual Description,* the *Annotator* step was used (see Figure 50). It augments the semantic expressivity level of triples with *low stamp*. The mapping file used as input in the *Annotator* step is illustrated in Figure 51.

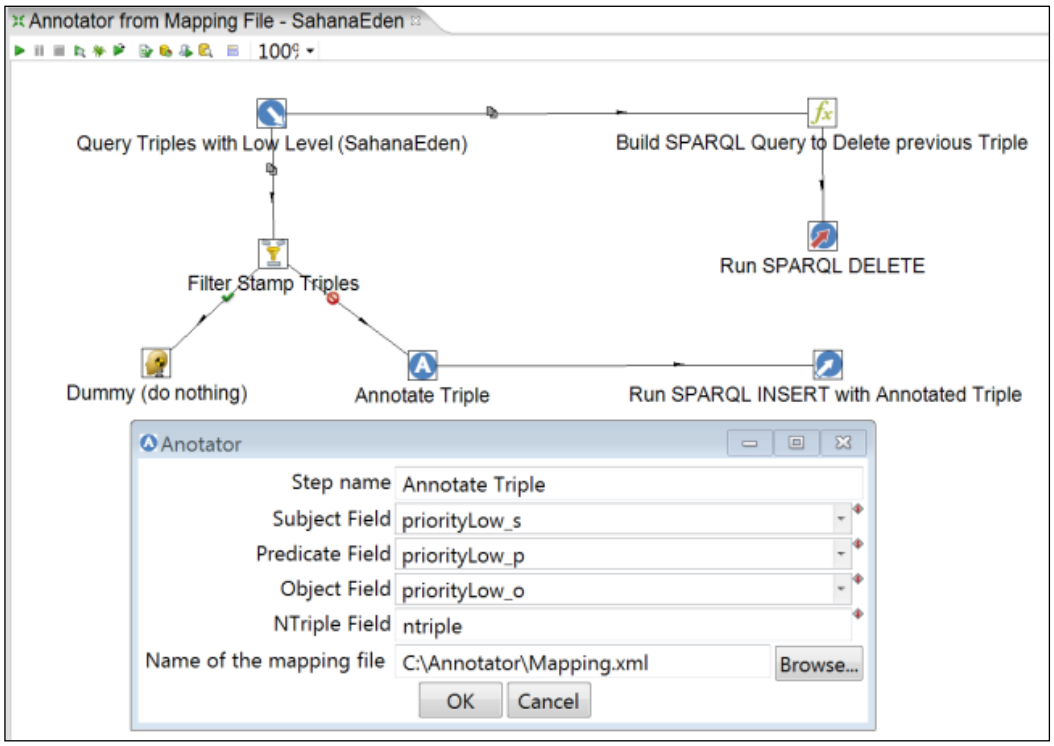


Figure 50: *Annotator* step to improve the semantic expressivity of Sahana Eden triples

```
<map id="1">
   <from>ToWarehouse/Facility/Office</from>
   <to>laiid:sentTo</to>
</map>
<map id="2">
   <from>DateSentOut</from>
   <to>time:hasDateTimeDescription</to>
</map>
<map id="3">
   <from>Name</from>
   <to>rdfs:label</to>
</map>
<map id="4">
   <from>City/Municipality</from>
   <to>place:Municipality</to>
</map>
```

Figure 51: Subset of the mapping file of the application case

The *Annotate with Conceptual Description* activity modified the previous triple set. Thereafter, the *Semantic Level Identifier package* is executed again. A subset of the annotated triples with the new stamps is shown in Figure 52.

<http://laiid.org.br/sahanaeden/resource#shipment254><http://www.w3.org/1999/02/22-rdf-syntax-ns#**type**><http://laiid.org.br/ontology/emergency/**Shipment**>
<http://laiid.org.br/sahanaeden/resource#shipment254><http://laiid.org.br/ontology/emergency#sentTo>"FO7 WAREHOUSE"
<http://laiid.org.br/sahanaeden/resource#shipment254><sstamp:hasSemanticLevel><sstamp:high>

<http://laiid.org.br/sahanaeden/resource#warehouse26><http://www.w3.org/1999/02/22-rdf-syntax-ns#**type**><http://laiid.org.br/ontology/emergency/**Warehouse**>
<http://laiid.org.br/sahanaeden/resource# warehouse26><http://laiid.org.br/ontology/emergency#name>"FO7 WAREHOUSE"
<http://laiid.org.br/sahanaeden/warehouse26><http://www.w3.org/2003/01/geo/wgs84_pos#**location**>
<http://laiid.org.br/sahanaeden/resource#**Region_VII**>
<http://laiid.org.br/sahanaeden/resource#warehouse26><sstamp:hasSemanticLevel><sstamp:high>

<http://laiid.org.br/ocha/resource#priority101><http://www.w3.org/1999/02/22-rdf-syntax-ns#**type**><http://laiid.org.br/ontology/emergency/**BeneficiaryGroup**>
<http://laiid.org.br/ocha/resource#priority101><http://laiid.org.br/ontology/emergency#**BeneficiaryQuantity**>
"74785"^^http://www.w3.org/2001/XMLSchema#integer>
<http://laiid.org.br/ocha/resource#priority101><http://www.w3.org/2003/01/geo/wgs84_pos#**location**>
<http://laiid.org.br/ocha/resource#**Region_VII**>
<http://laiid.org.br/ocha/resource#priority101><sstamp:hasSemanticLevel><sstamp:high>

Figure 52: Annotated triples marked with *high stamp triples*

After processing the data sources, doing some cleaning, conforming transformation, converting to triples, marking their semantic level, and augmenting the semantic expressivity, a set of RDF Graphs and *stamp triples* were generated, as described in Table 6. Note that the number of semantic level *stamp triples* is the same as that of RDF Graphs, which group the triples in Subject Items.

Table 6: Profile of data sources transformation output

Source	Application Case Scope	Data Demo Scope	Domain Triples	RDF Graphs	Number of Semantic Level Stamps
Sahana Eden	Shipment	2045 rows 3 columns	2184*	728	728
	Warehouse	48 rows 4 columns	240	48	48
OCHA/DSWD	Priority	171 rows 5 columns	1026	171	171

*Filter applied: field.sentTo not null

The last activity is the *Adaptive Interlinking*. This means that an appropriate interlinking approach was defined for a dataset with the same semantic level. Thus, the interlinking approach, which depends on schema level descriptions, was used to create the links between triples with a high semantic level. In this case, as explored in Section 3.5, the Silk Server was used. Figure 53 shows the implemented job.

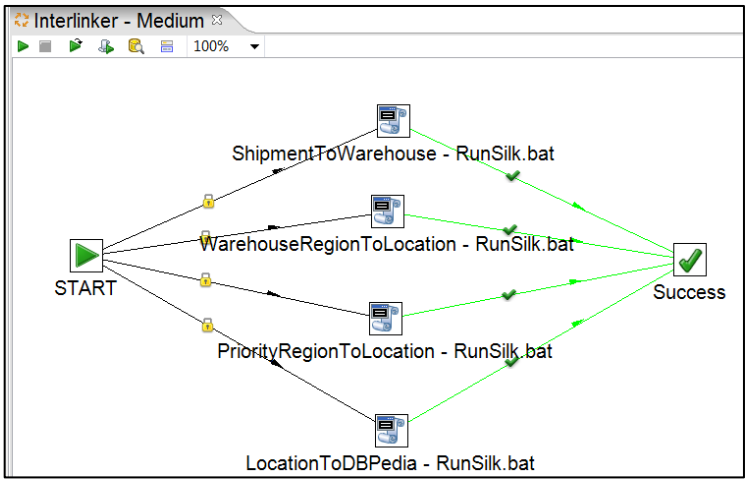


Figure 53: Job of Silk interlinking steps

On the other hand, the interlinking approach, which only uses literal object matching, was applied to create links among triples with a lack of semantic concerns. In this case, the step Object Property Mapping of the ETL4LOD was used (see Figure 54 and Figure 55).

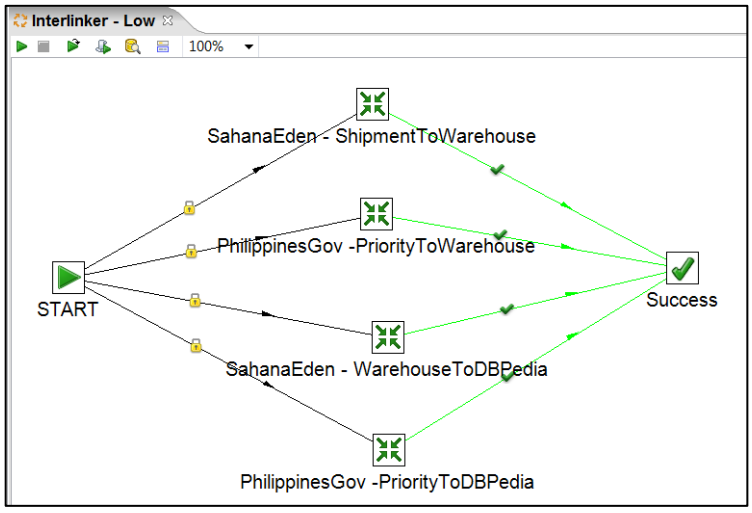


Figure 54: Job of the set of ETL4LOD transformations

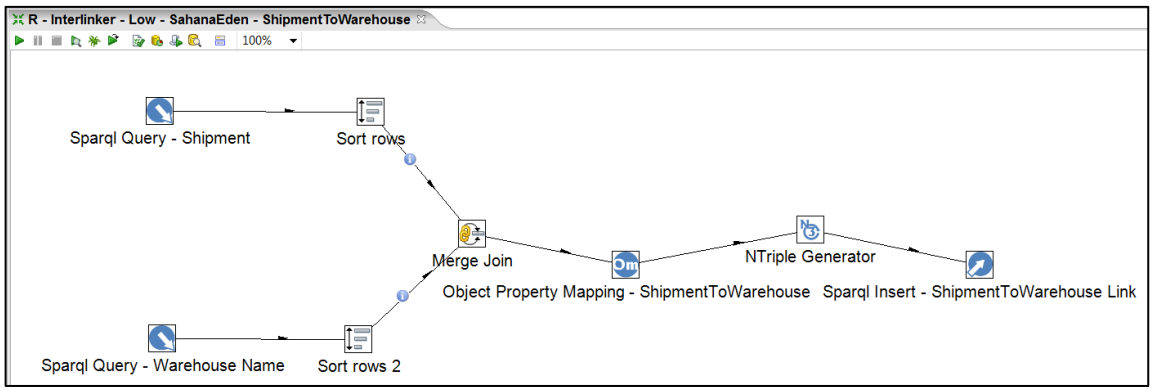
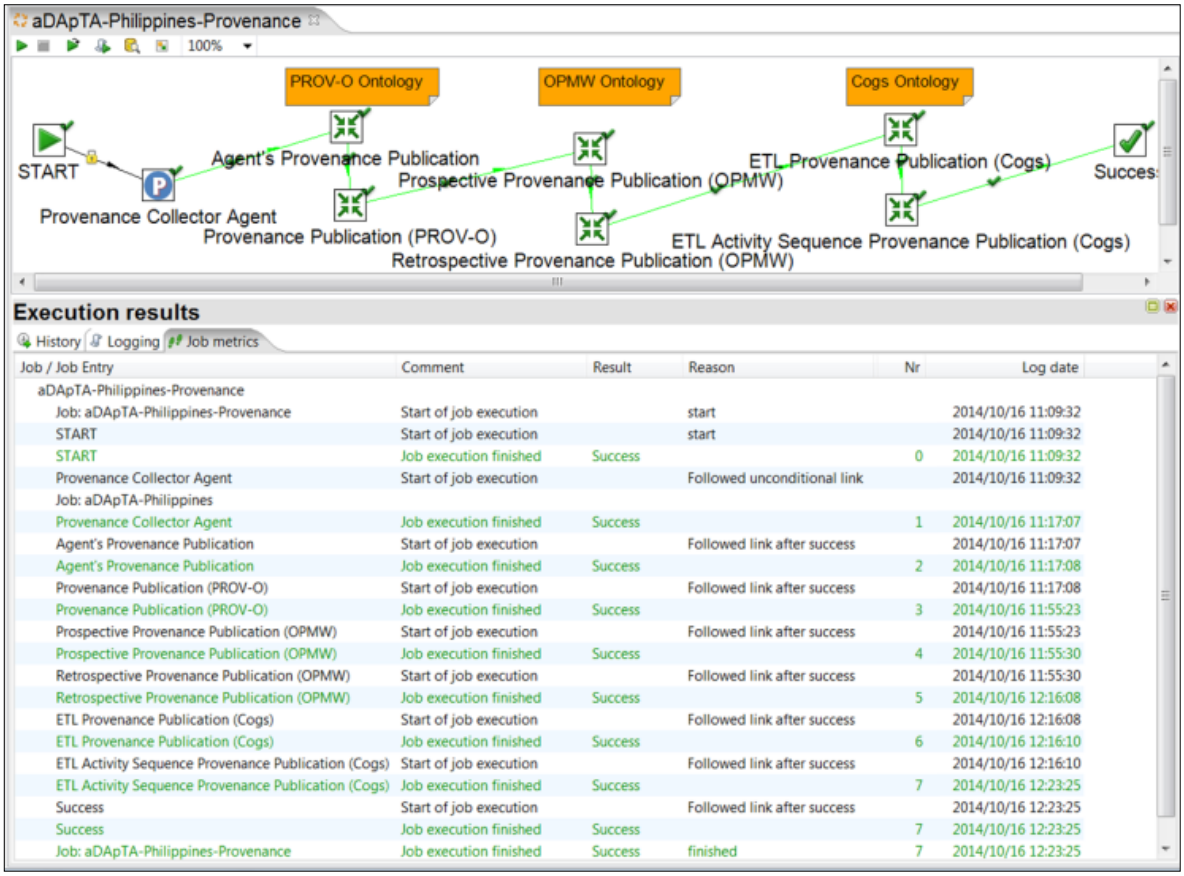


Figure 55: Interlinking package of the Sahana Eden data marked with *low stamp triples*

5.4 Provenance Collector Agent

In parallel to the *Adaptive Integration of Information* activities, to collect the provenance data of the ETL workflow, a *Provenance Collection Agent* was added to the workflow. The agent was configured to collect the provenance of each activity of the approach, implemented as a sub-workflow. After collection, the provenance data were published with the semantic support of some ontology and interlinked to the domain data. Figure 56 shows the implemented workflow and its execution log.



Job / Job Entry	Comment	Result	Reason	Nr	Log date
aDApTA-Philippines-Provenance					
Job: aDApTA-Philippines-Provenance	Start of job execution		start		2014/10/16 11:09:32
START	Start of job execution		start		2014/10/16 11:09:32
START	Job execution finished	Success		0	2014/10/16 11:09:32
Provenance Collector Agent	Start of job execution		Followed unconditional link		2014/10/16 11:09:32
Job: aDApTA-Philippines					
Provenance Collector Agent	Job execution finished	Success		1	2014/10/16 11:17:07
Agent's Provenance Publication	Start of job execution		Followed link after success		2014/10/16 11:17:07
Agent's Provenance Publication	Job execution finished	Success		2	2014/10/16 11:17:08
Provenance Publication (PROV-O)	Start of job execution		Followed link after success		2014/10/16 11:17:08
Provenance Publication (PROV-O)	Job execution finished	Success		3	2014/10/16 11:55:23
Prospective Provenance Publication (OPMW)	Start of job execution		Followed link after success		2014/10/16 11:55:23
Prospective Provenance Publication (OPMW)	Job execution finished	Success		4	2014/10/16 11:55:30
Retrospective Provenance Publication (OPMW)	Start of job execution		Followed link after success		2014/10/16 11:55:30
Retrospective Provenance Publication (OPMW)	Job execution finished	Success		5	2014/10/16 12:16:08
ETL Provenance Publication (Cogs)	Start of job execution		Followed link after success		2014/10/16 12:16:08
ETL Provenance Publication (Cogs)	Job execution finished	Success		6	2014/10/16 12:16:10
ETL Activity Sequence Provenance Publication (Cogs)	Start of job execution		Followed link after success		2014/10/16 12:16:10
ETL Activity Sequence Provenance Publication (Cogs)	Job execution finished	Success		7	2014/10/16 12:23:25
Success	Start of job execution		Followed link after success		2014/10/16 12:23:25
Success	Job execution finished	Success		7	2014/10/16 12:23:25
Job: aDApTA-Philippines-Provenance	Job execution finished	Success	finished	7	2014/10/16 12:23:25

Figure 56: Execution of the *Provenance Agent* applied to the case

Performing the process detailed in Section 4.2 (Provenance Collection, Publish, and Interlink), more than two million provenance triples were generated to support the quality assessment of the integrated process of data described in Table 6. The finest grain was used. This means that the retrospective and prospective provenance of all step types of the workflow was collected, published and interlinked. Despite the great volume, it took only 1 h, 14 m, 9 s to perform the whole process in a computer with an Intel i7 processor, 8GB of

memory and an SSD hard disk. The triple store was also hosted on a remote server implemented in a Virtual Machine with two processors and 4GB of memory.

5.5 Decision Making Support

With data interlinked from different sources and provenance data collected, some reports were generated to support decision making. Sgvizeler[32] was used to create online reports over the aDApTA SPARQL endpoint available at http://crbd.ppgi.ufrj.br:8890/sparql.

In the scenario case, one of the first information demands to support decision making focuses on the affected people. Figure 57 shows a hierarchic graph of the regions, cities, and provinces of the Philippines and the number of displaced people represented by the size of the circles. Thus, the report indicates that the province Leyte of Region VIII and province Cebu of Region IV-B demands urgent humanitarian aid.

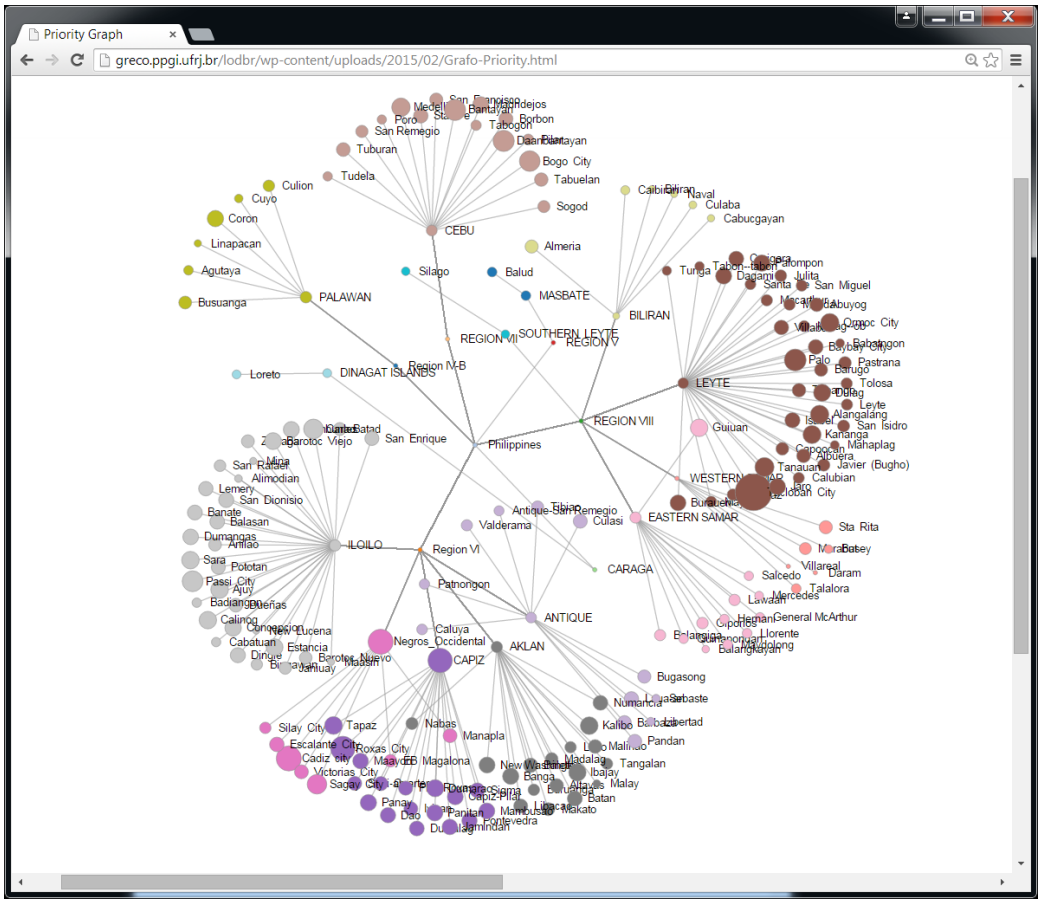


Figure 57: Graph of an integrated view for decision support

[32] http://www.w3.org/2001/sw/wiki/Sgvizler

With this report, the decision maker may ask which data sources were used and who are the corresponding data owners. This information can be retrieved from the provenance repository. The performed SPARQL query and corresponding result are displayed in Figure 58.

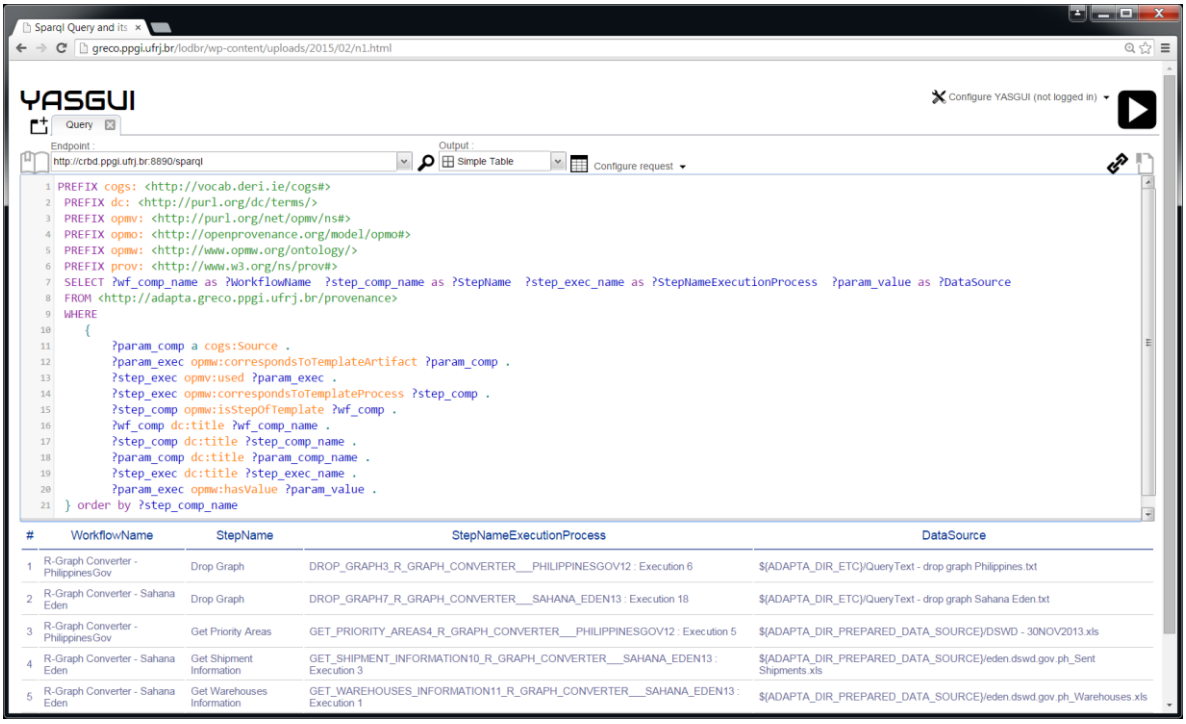


Figure 58: SPARQL query and its results for provenance data

The provenance data reveals that the owner of the dataset is the OCHA/DSWD, a reliable source. Subsequently, the decision maker may ask for the population of a specific area. The treemap illustrated in Figure 59 shows the relation between the province of Leyte and the population versus displaced people.

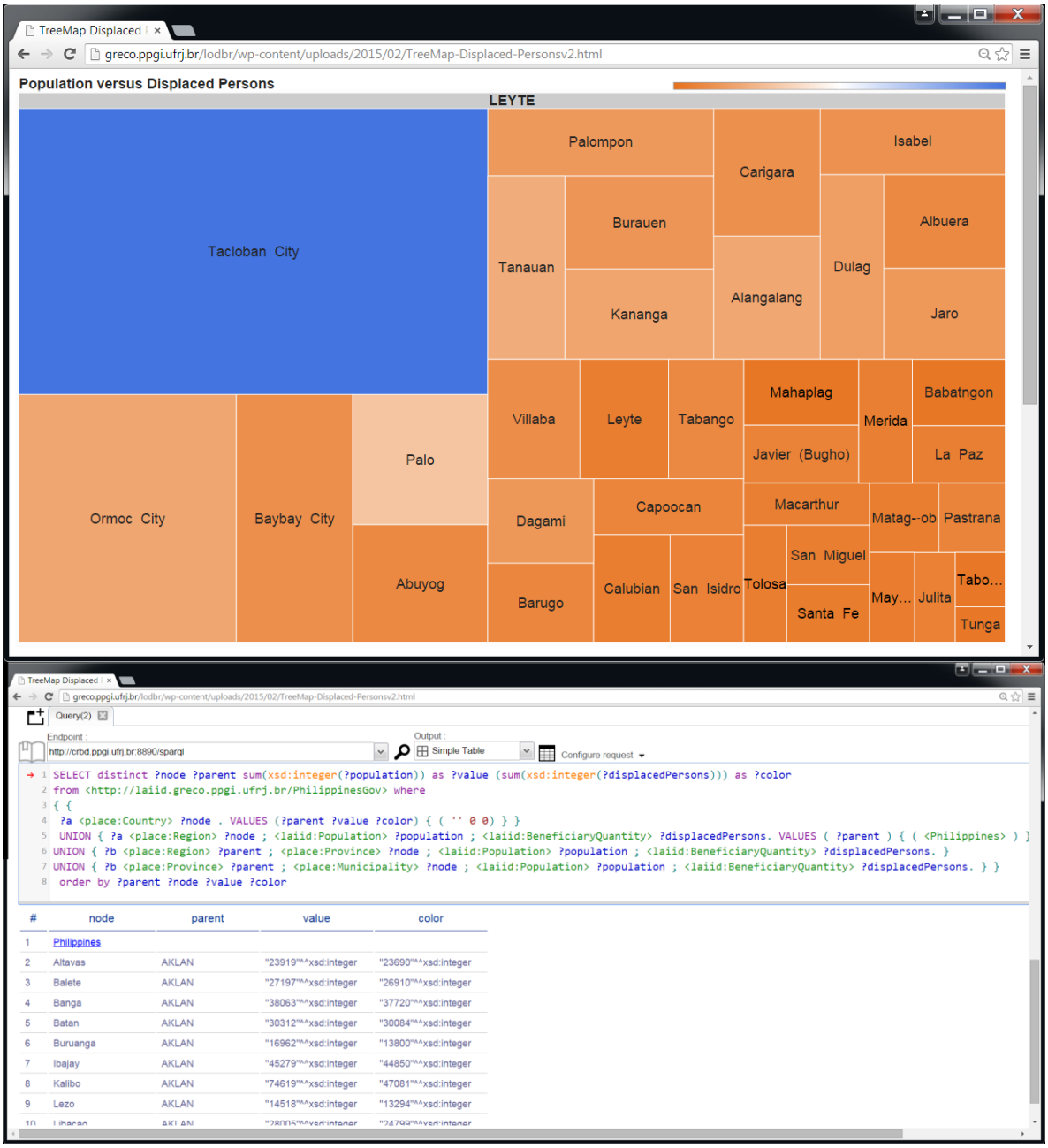


Figure 59: Relation between the *Population* versus *Displaced People* of Leyte province. Report and SPARQL query interface

After one has an idea of the number of affected people, the geo-location should be analyzed. In the Philippines application case, it is of special importance to take into account the characteristics of the geographic region, an archipelago of more than 7000 islands. The complementary information about the latitude and longitude coordinates was retrieved through a federated SPARQL query on the aDApTA endpoint and the DBPedia endpoint. The circle size and color strength, displayed in Figure 60, represent the number of displaced

people. The information on the location, the percentage of displaced persons, and a link to DBPedia with the information about the location were used to build the map illustrated in Figure 61, which shows a map of the Philippines with the affected locations marked with red pins. Moreover, it is possible to view the same information in the tabular sheet and to analyze the map from different granularities, selecting Region, Province, or Municipality perspective.

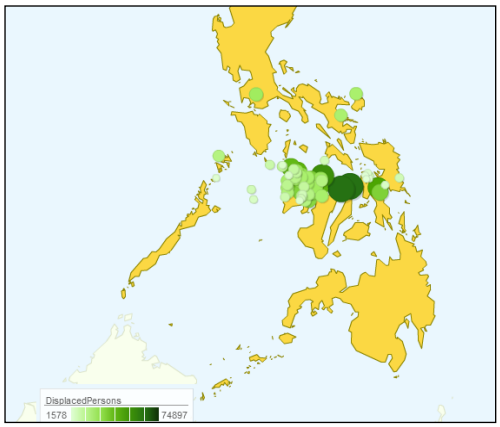


Figure 60: Integrated view for decision support in a geomap

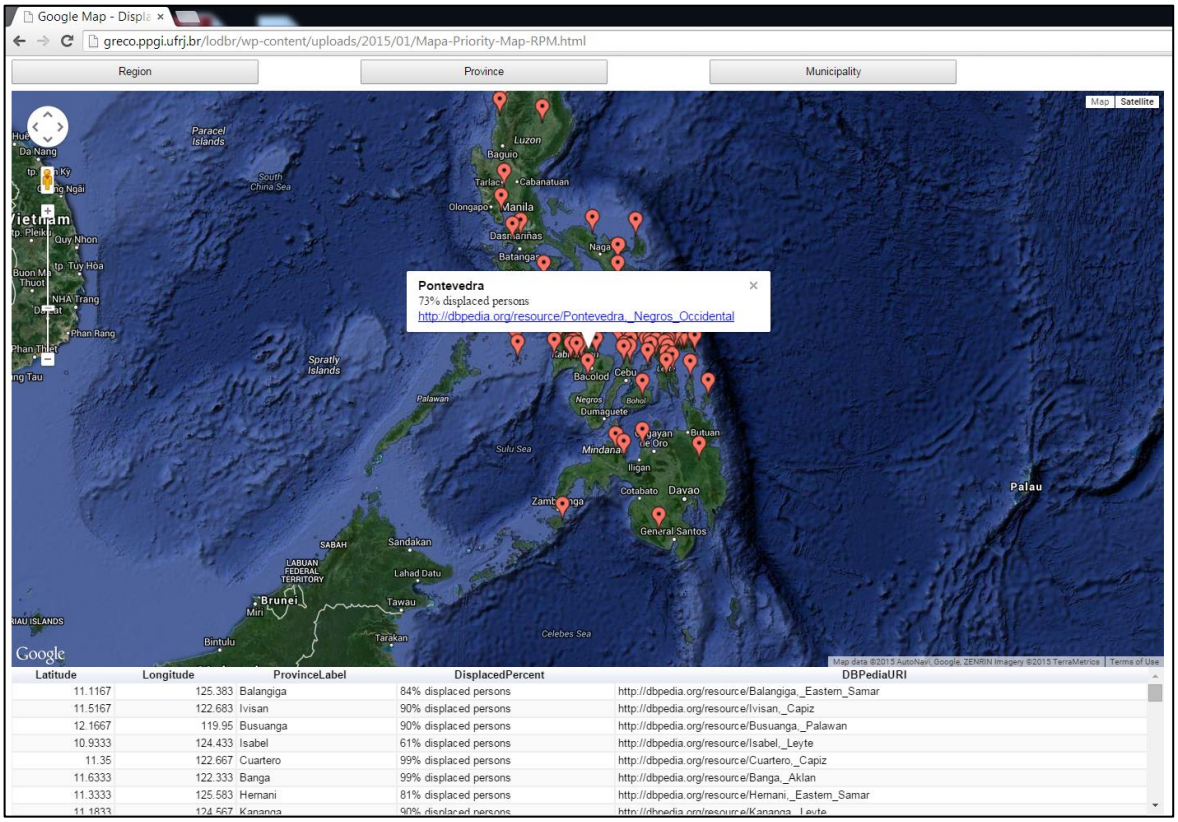
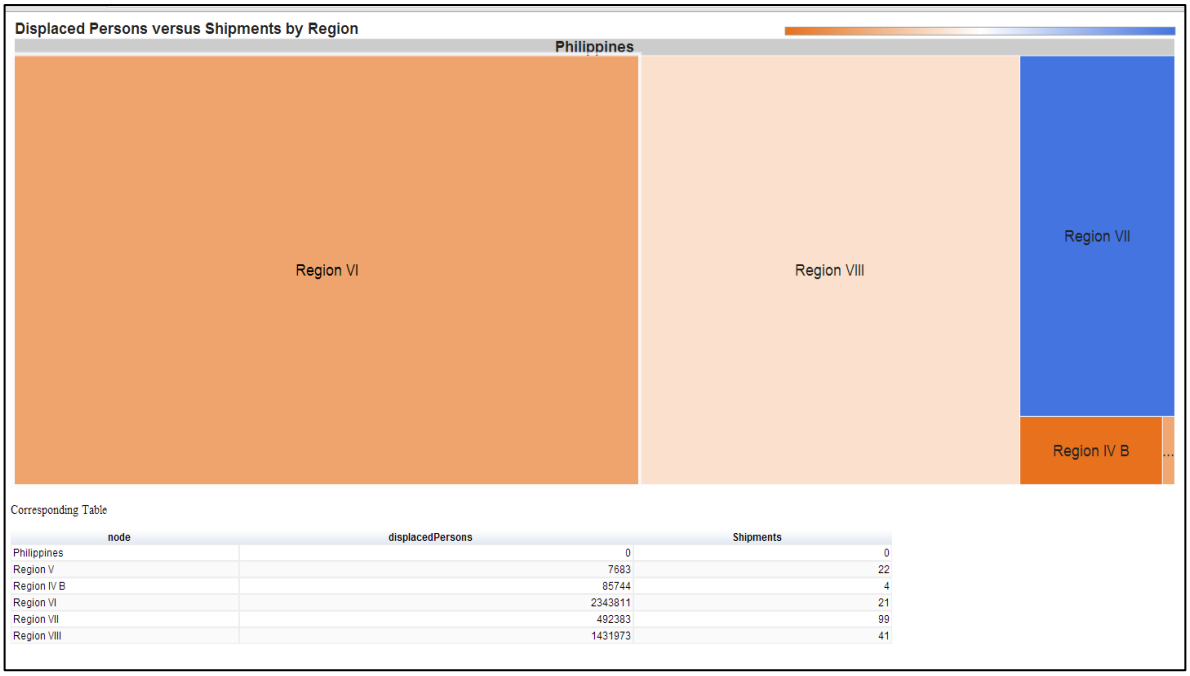


Figure 61: Integrated view for decision support in a GeoMap exhibiting percentage of *Displaced People* from *Region*, *Province,* or *Municipality* perspective

Additionally, the integrated view reported in Figure 62 can be used to monitor the correlation of the shipment numbers and displaced people by Philippine regions. The treemap rectangle size represents the amount of displaced people, and the color strength represents the number of shipments sent to the warehouses of that region. For example, Region VII has more displaced persons than Region IV B, and many more shipments.



node	displacedPersons	Shipments
Philippines	0	0
Region V	7683	22
Region IV B	85744	4
Region VI	2343811	21
Region VII	492383	99
Region VIII	1431973	41

Figure 62: Integrated view for decision support in a TreeMap: *Displaced People* versus *Shipments* by Region

Information about the population and of the displaced people on each island is the basis for the decision of which boats are going to be used to ship relief goods to which set of islands. To make the final decision about the priority areas for the shipments, ultimate information was requested. When was the last update of data? The requested information was retrieved from the provenance repository again. Figure 64 shows a screen shot of a Web report about the creation and modification of each ETL package. Moreover, Figure 64 shows a screen shot of a Web report about the execution timestamp of each ETL package. In addition, it shows if the job was successfully completed. Based on that, the decision maker became aware that the data from the Philippine Government was loaded on November 16, 2013.

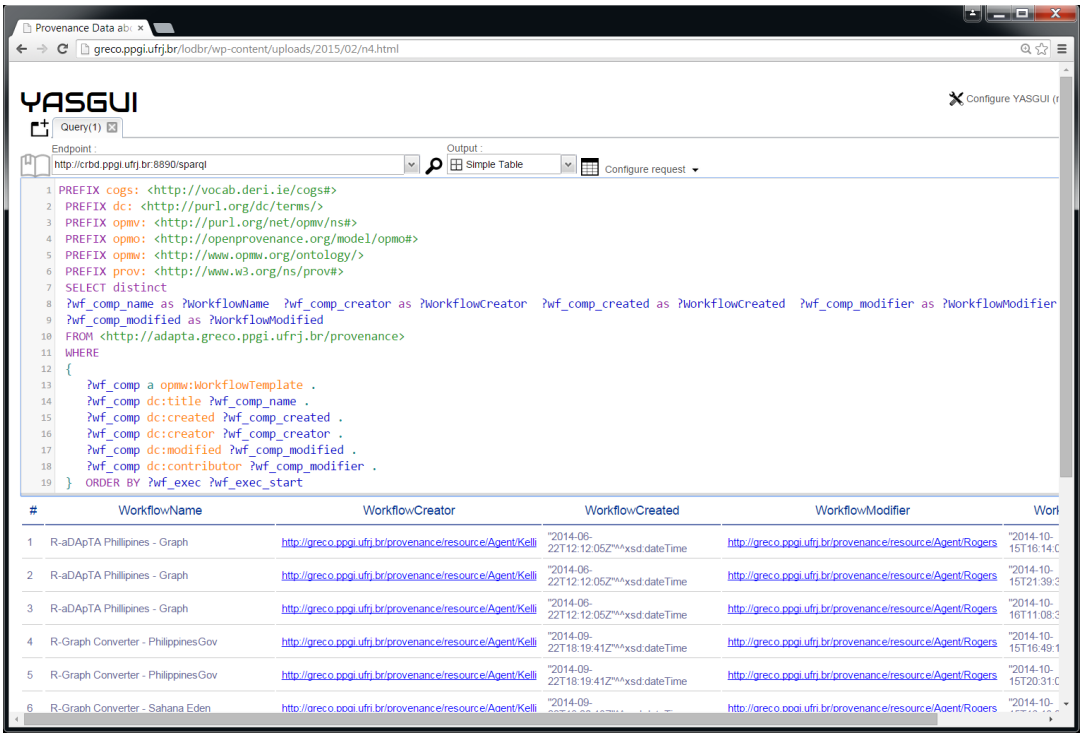


Figure 63: Provenance data about the ETL package update

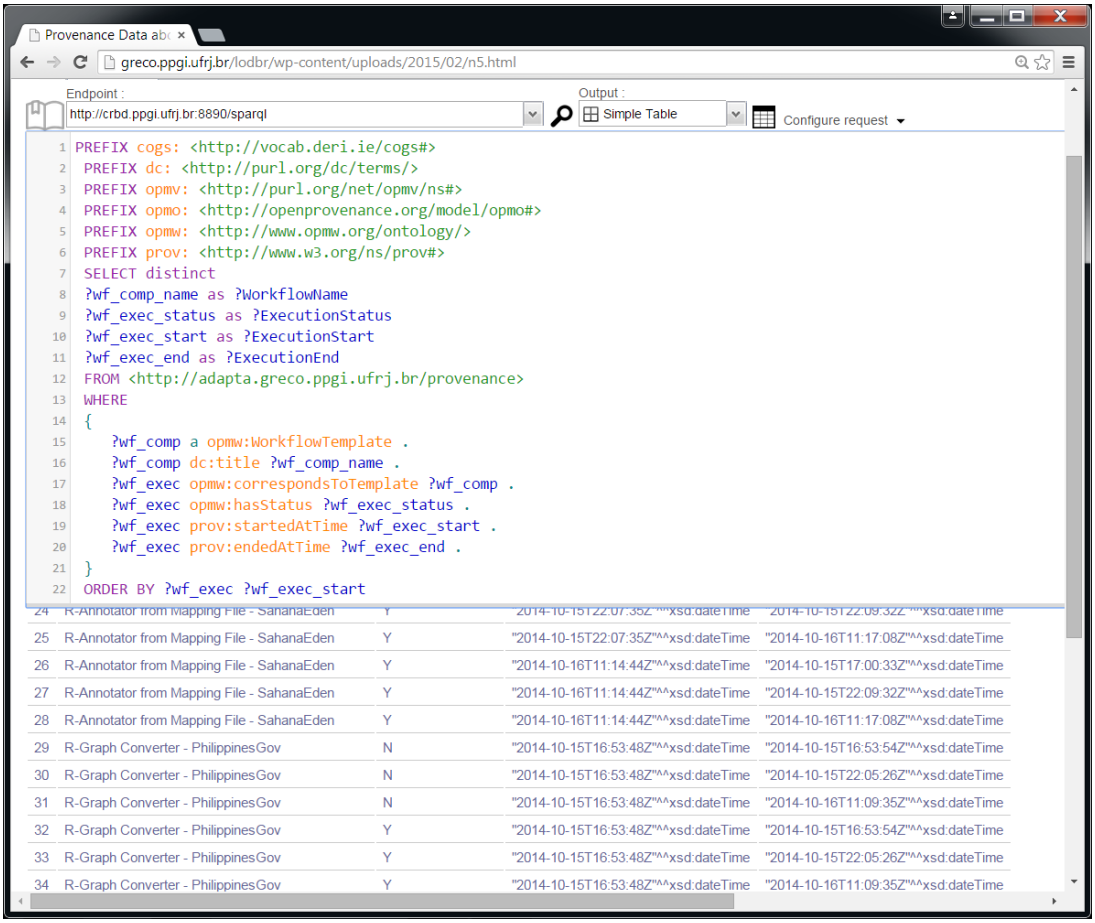


Figure 64: Provenance data about the ETL package execution

Besides the integrated view of the crisis for the decision makers, it is possible to develop views about the data profile to support the information analyst to make some improvements to the system. The number of triples, properties, classes, and objects can be compared using the graphics presented in Figure 65.

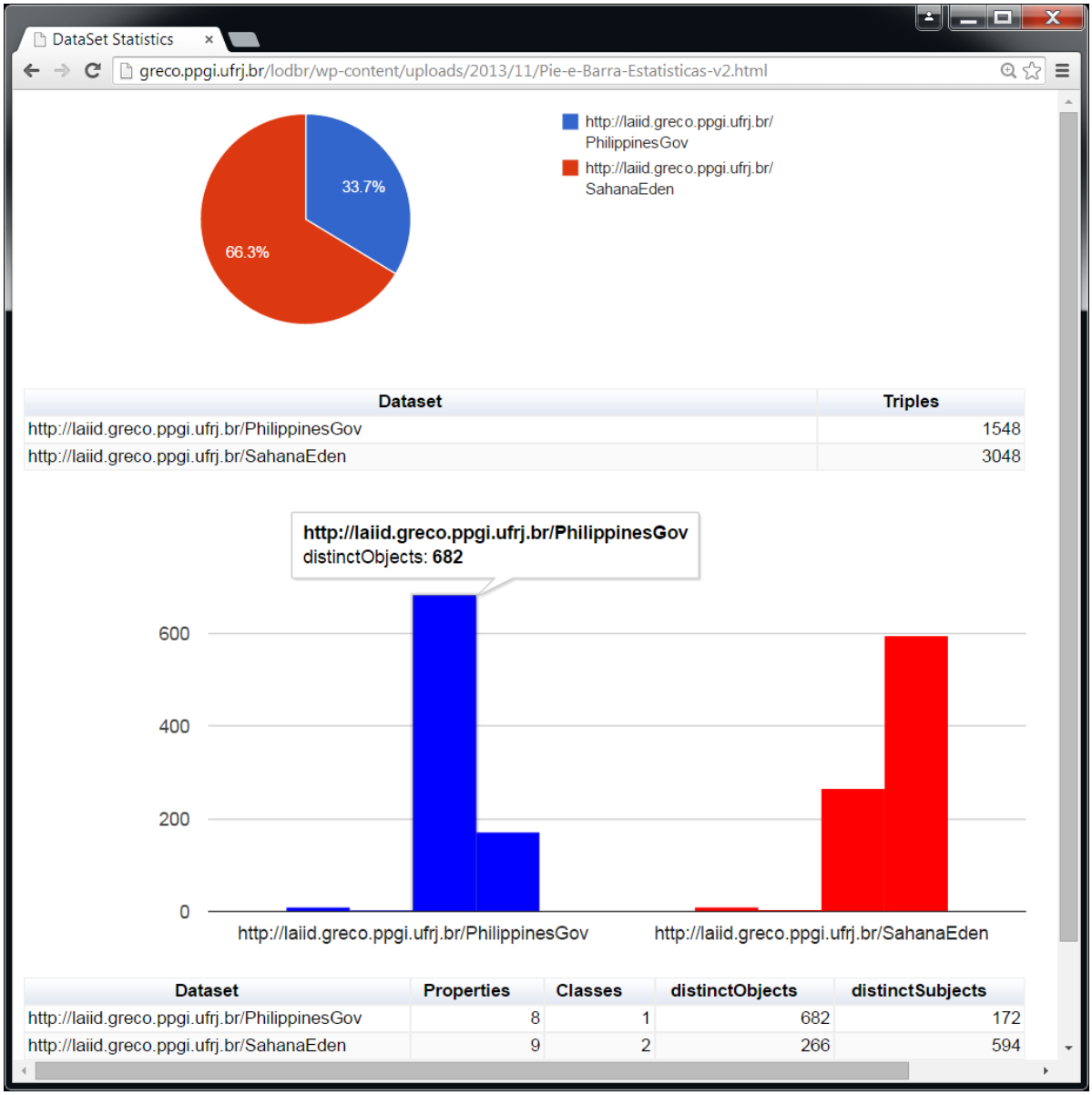


Figure 65: Interlinked data profile

One of the main contributions of this work is partially illustrated in the following figures. The semantic expressivity level of the Philippines Government's data source about the Priority regions was evaluated as low. Thus, it was marked with *low stamp triples* (see Figure 66). The semantic level of the Sahana Eden data source about the Shipments and Warehouses was evaluated as high. Hence, it was marked with *high stamp triples* (see Figure 67).

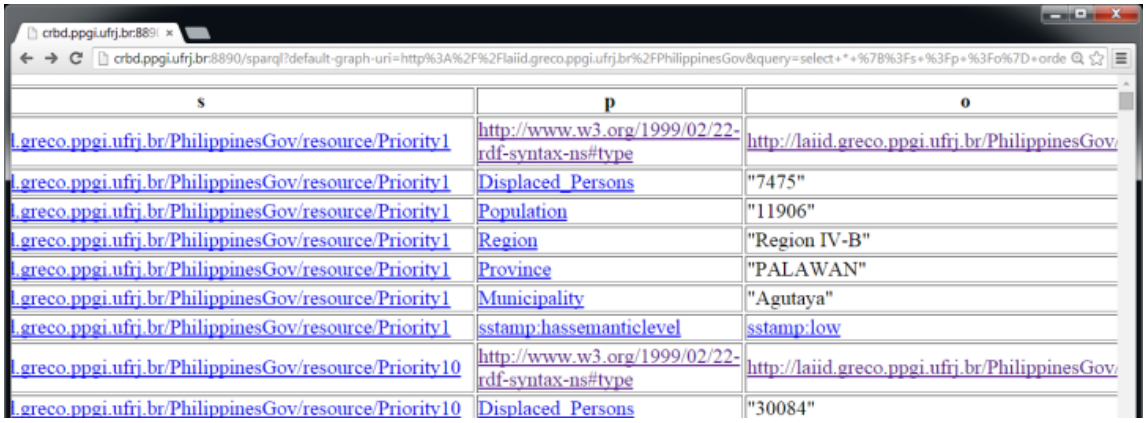


Figure 66: Data soure marked with *low stamp triple*

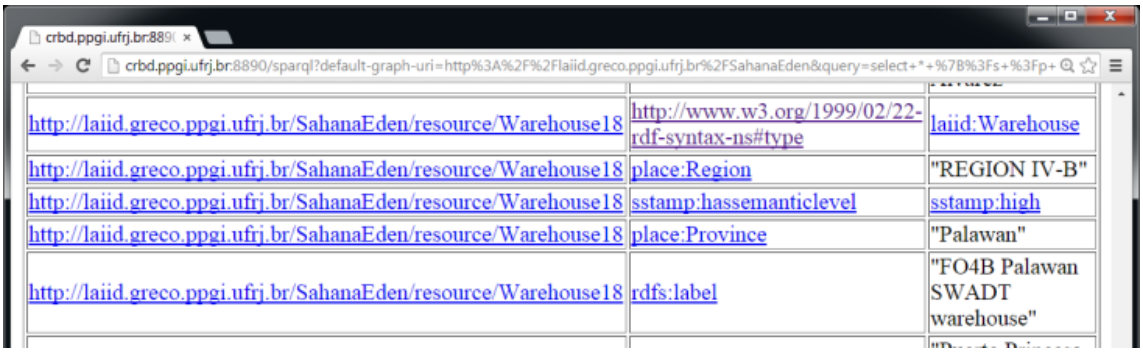


Figure 67: Data soure marked with *high stamp triple*

There is a clear semantic expressivity level conflict between the data sources. One data source is marked with *high*, as it uses concepts of a ontology to describe its subjects, and the other with *low* semantic expressivity level, as it only use literal values to describe its subjects.

In order to adapt the system, the conflict was solved by performing some steps. The first one was running the *Annotator* step to improve the semantic level of Priority regions. The second one was the definition of the Silk Server to create the links. It was performed by selecting the interlinking approach also marked with *high stamp triple.*

The interlinking approach used to integrate the data sources marked with *high stamp triples* is displayed in Figure 68, which illustrates the screen shot of a SPARQL query and the corresponding result, the Silk Server. In addition, the provenance data about the ontologies used to improve the semantic level of the Priority data source is shown in Figure 69.

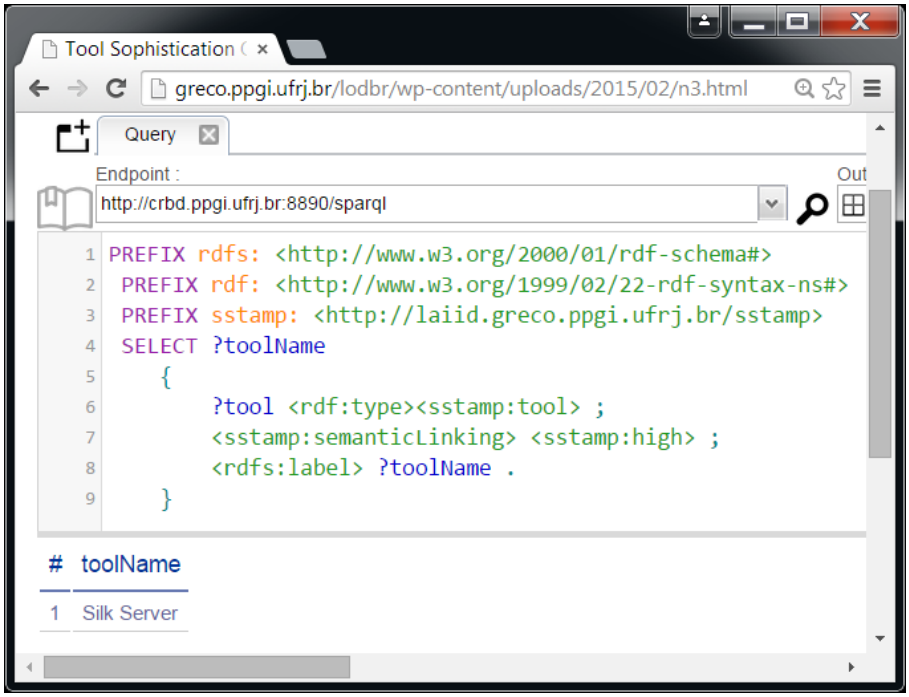


Figure 68: SPARQL query of the interlink tool marked with *high stamp triple* and the result: the Silk Server.

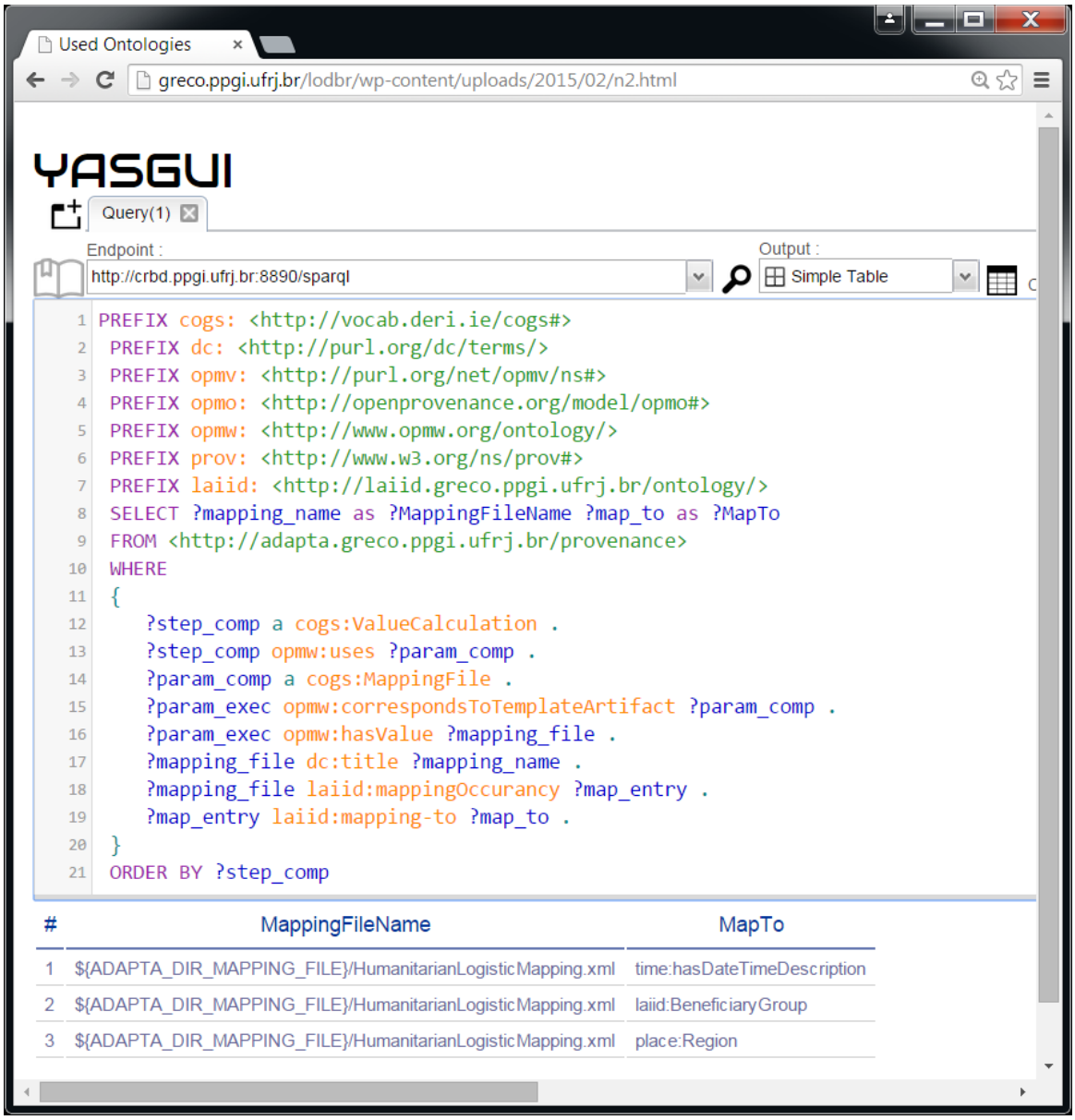


Figure 69: SPARQL query of the used ontologies and the result: time, laiid and place.

5.6 Approach Evaluation and Discussion

Complex environments are not trivial to be simulated in a laboratory to evaluate a solution approach. Thus, data from a real event were gathered in order to demonstrate the feasibility of aDApTA. The application case illustrated the issues of integration information under dynamic environment and the corresponding adaptive solutions. The information analyst is the one who can have the better perception of the aDApTA results, as this role is responsible for delivering the information demands. In this scenario, the approach was evaluated through the analyses of each activity.

The result of the graph data conversion and triplification process was up to four thousand triples, with hundreds of distinct subject and objects. Thereafter, few properties and classes were used in the annotation and interlinking process, showing that the resulting dataset is composed of many instances of few entities. This information is presented to the information analyst through a dashboard report based on SPARQL queries result, as presented in Figure 70.

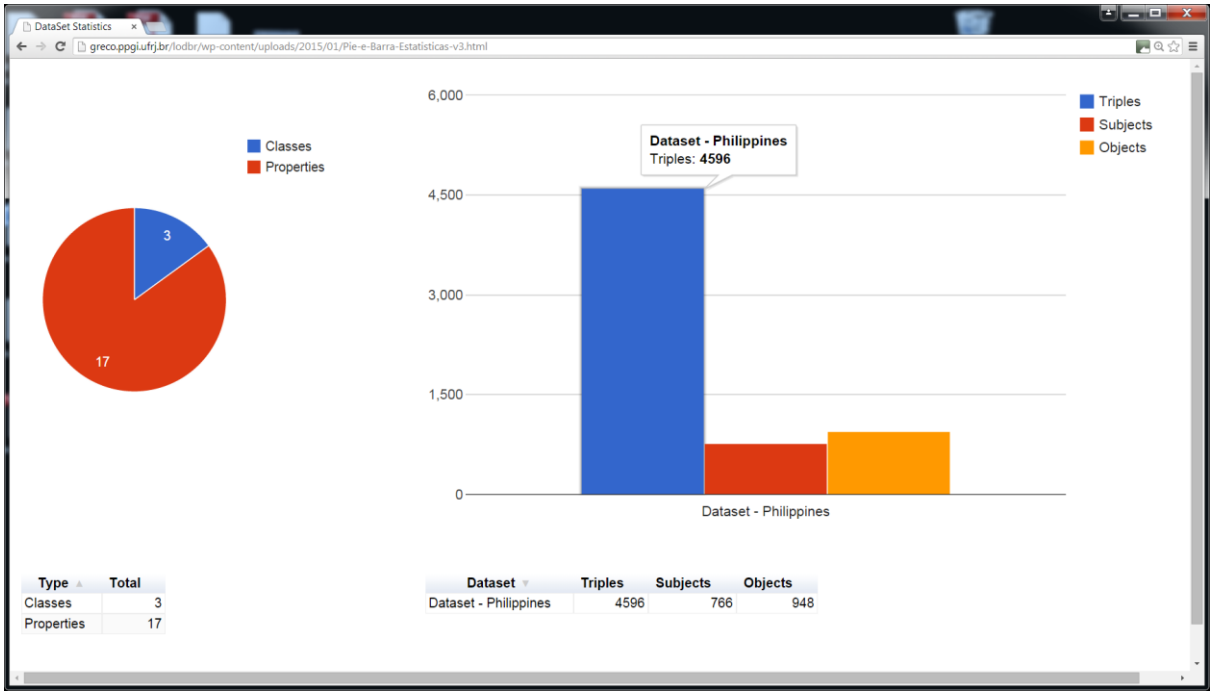


Figure 70: Dataset analyses.

Another perspective to analyze the data is the semantic level assessment results showing a low expressivity level of data sources. Almost one thousand triples were marked with *low stamp*. Within this scenario, the interlinking approach performed by a tool that only used literal values to create the links, generated almost two hundred triples of external links, as presented in Table 7.

Table 7: Interlinking results

		Stamp Triples	
		Low	High
Link Type	Internal	757	747
	External	187	12
Total		944	759

However, after annotating the data with concepts from the domain ontology, marking them with *high stamp*, and choosing a tool that uses the ontologies concepts to create the links, only twelve external links were generated. This metric was used to indicate the semantic improvement achieved on the integrated data by adapting the interlinking approach to the semantic expressivity level of data sources.

In addition to this perspective, after the application of the interlinking tools, and the differentiation of the semantic levels, the number and type of the generated links can be analyzed through the graphic presented in Figure 71.

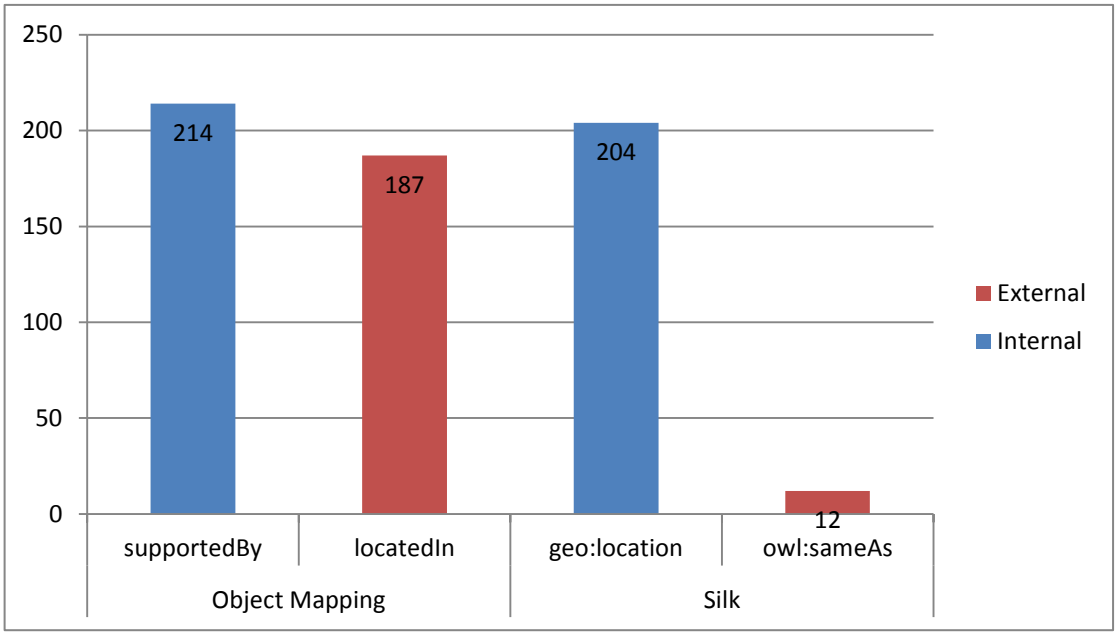


Figure 71: Profile of the interlinking result set

With more semantic expressivity, fewer links are created. In the application case, it happens because of the use of domain ontology. Figure 72a depicts the links generated among triples from different data sources marked with *low stamp triples*, and Figure 72b depicts the links after the annotation step to improve its semantic expressivity. It is then marked with *high stamp triple*. Note that the *Warehouse* and *Place* entities of DBPedia are connected with direct links when the entities are marked with *low stamp triple*. However, when they are marked with *high stamp triple*, they are connected through the *Region* entity of the domain ontology.

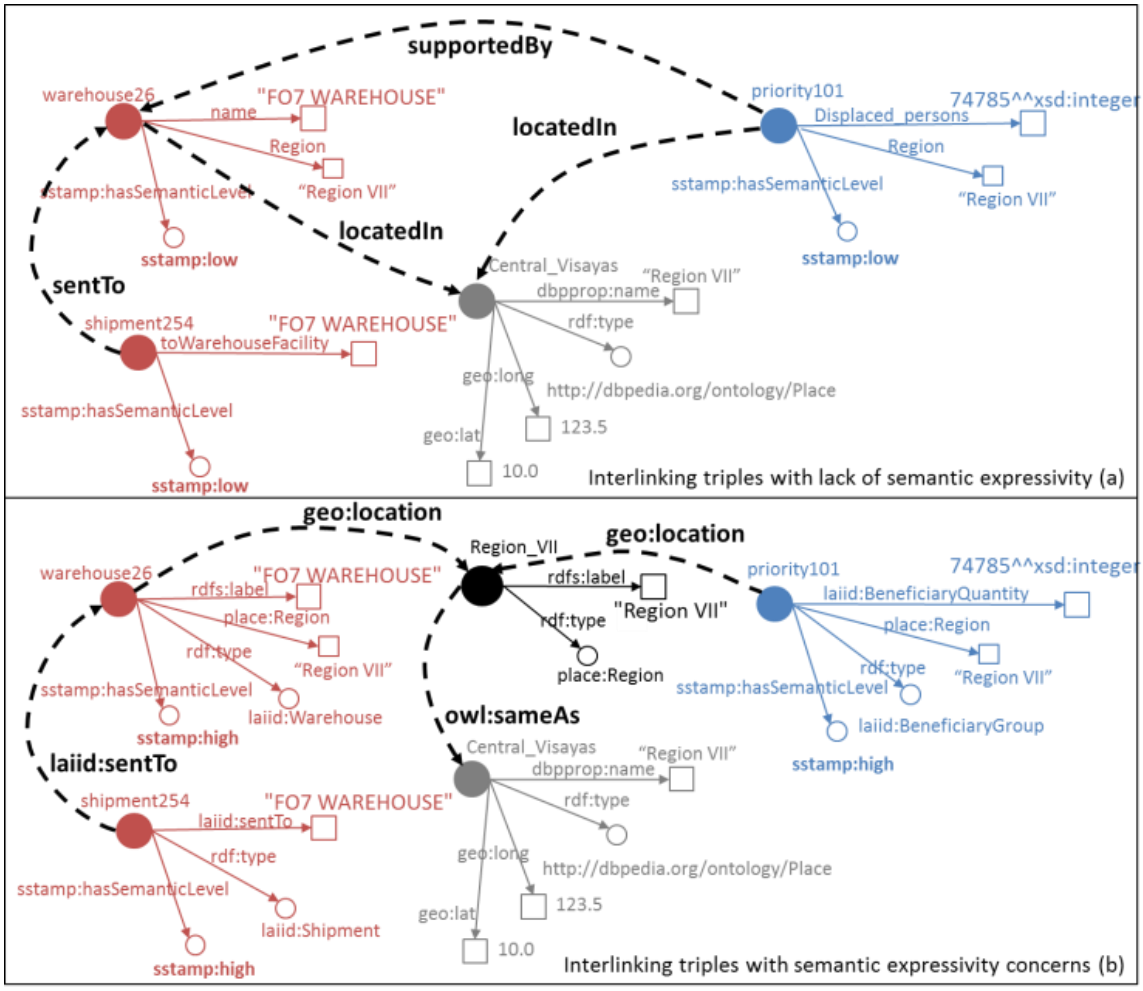


Figure 72: Interlinking triples with different semantic expressivity levels

Thus, the external links decrease from 204 to only 12, indicating how better links can be created as more expressive the data are. Figure 73 and Figure 74 depict the number and type of links between data sources with low semantic expressivity in contrast to data sources linked through domain ontology concepts.

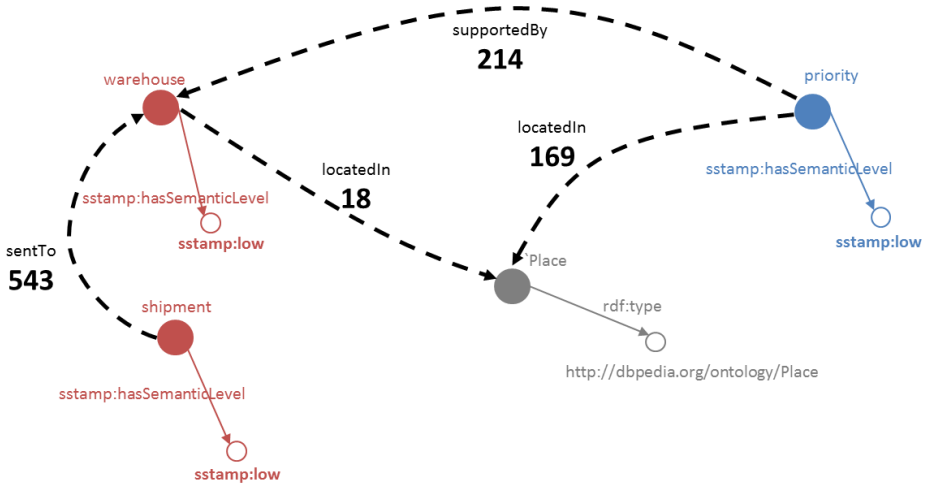


Figure 73: Number and types of links between data sources with low semantic expressiveness

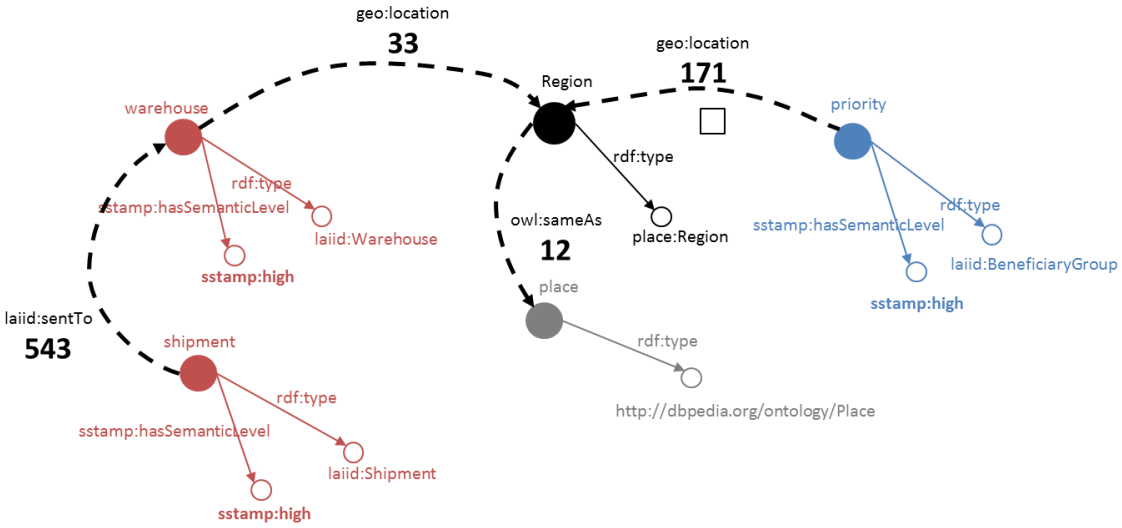


Figure 74: Number and type of links between data sources with high semantic expressiveness

Based on these results, the evidence suggests that the more expressive the data are, the better links are created, i.e., fewer and more expressive ones. With better links, more optimized reports can be built. Moreover, it reduces the demand for a computational resource from others interlinks jobs by minimizing the redundancies on links. Another important aspect is the link type. Using common vocabularies and ontologies facilitates the

reuse of data from third party applications. In the application case, the RDF Graphs marked with *high stamp triples* use common vocabularies.

Finally, the approach evaluation highlighted one of the main contributions of this thesis, the application of a stamp to mark the semantic expressivity level of an information resource. The stamps allow the choice of a more appropriate interlinking approach, adapting the information integration system to unpredicted data sources. However, some design decisions must be considered while implementing this feature of aDApTA. Some examples are: stamp only what is required for a given application case; stamp different levels of granularity, such as graph, sub-graph, dataset or a single triple; stamp based of graph metrics, such as closeness and betweenness centrality; the semantic expressivity evolves as new links are generated, the stamps must be changed with it.

The evaluation of the state of the art and the alternative solutions to handle information integration in a complex environment were explored in Section 2.3 (Open Issues in Information Integration in Complex Environments) highlighting open issues (see Table 1). Some complexity aspects were addressed: the information semantic expressivity heterogeneity, the information trustworthy, and the unpredictability of the data sources characteristics. The aDApTA handles these issues using an adaptation method by evaluate the emerge heterogeneity and switching the integration component accordingly. To handle the trustworthiness, a critical issue in decision making, the provenance data were used. Also, some supporting tools were listed in Section 3.5 (Supporting Tools, Frameworks and Architectures for Linking Data) showing gaps that were addressed by the ETL4LOD-Graph steps (see Table 3). The aDApTA evaluation described in this chapter detailed the approach. The implementation details and source code links are explored in Appendix B.

6 Conclusions

This research work described aDApTA, an adaptive approach for information integration to support decision making in complex environments. aDApTA uses Linked Open Data principles as a strategy to support the adaptation of the integration approach solving data structure and description conflicts. The main strategy is to match the semantic expressivity level of data sources with the sophistication level of the interlinking methods. The results indicate that links can be generated when the integration approach is adapted to the characteristics of the incoming data, even though they cannot be predicted. aDApTA enables the creation of links by resolving semantic and structural conflicts between data sources. Moreover, it minimizes the redundancies and supports the reuse of integrated data by using domain ontologies and common vocabularies. The application case prototype showed the feasibility of aDApTA and that it can meet the requirements of a decision support system in complex environments. A study on how Web standards and the Web of Data can support adaptive information bases was presented, as well as an analysis of how integration of heterogeneous data can be achieved by using graph data representations. Finally, an easy-to-use open source framework was developed and made available, so that it can be applied to other domains.

6.1 Contributions

The solution described in this thesis started with the recognition of the semantic expressivity as a heterogeneity level to be handled by an information integration process, additionally to the syntax, structural, semantic, schema, identity and data levels. Thereafter an approach to solve this conflict was proposed based on a stamp. The approach evaluation highlighted the possibilities of using the stamps to mark the semantic expressivity level of an information resource. The stamps support the choice and allow the interlinking approaches switching accordingly to the semantic sophistication they consider. Thus, the triple stamps enable the adaptation of an information base by integrating data no matter the semantic expressivity it has. The evaluation focused on showing the improvements on integration

results, therefore, supporting decision making in complex environment with a unified view of the situation.

In summary, the contributions of this thesis are:

- Characterization of the dynamic of an information integration in a complex environment;
- Approach for adaptive information integration;
    - Supporting architecture;
- Semantic Level Evaluation Framework of an Information Resource;
    - Stamp triples concept; and
- Open Source ETL Framework that processes RDF graph data, additionally to other data formats.

Based on these contributions, the results of the thesis are:

- Prototype using data from a real case of decision support in complex environments;
- Evaluation of the approach through the prototype;

Furthermore, some outcomes of the research were:

- Presentation of the approach proposal in Cordeiro et al. (2014a) and publication in Cordeiro et al. (2014b); and
- Publication of aDApTA details and results in Cordeiro et al. (2015).

6.2 Limitations

The major problem addressed in this thesis is how to support the integration of information for decision making in a complex environment where most relevant data are heterogeneous and cannot be known in advance. Among the complex aspects, only heterogeneity and unpredictability were handled through an adaptation approach. The solution was limited to provide an approach to integrate heterogeneous data sources in an adaptive way. Another important complex aspect, feedback, stayed out of the thesis's scope, even though it is an important feature in dynamic environments, especially those with great interaction with end users. Uncertainty is also an important complex aspect present in a great volume of heterogeneous and distributed datasets where the queries are answered with assumptions.

In order to provide an integrated view, only the structural and semantic heterogeneity levels were processed. Others levels, such as syntax, identity and data were not handled; however, they could be treated by the chosen interlinking approach. In addition, unstructured data were not handled.

Complex environments are not trivial to be simulated in a laboratory to evaluate a solution approach. Retrieving data sources from a real scenario is also a difficult task as there is a lack of open systems and data sources. Thus, a subset of a dataset from a real event was gathered in order to demonstrate the feasibility of the approach. Based on this scenario, the scope of the prototype was restricted to show that choosing interlinking approaches accordingly to its level of semantic sophistication can improve the integration results. Concerning implementation, some issues were not handled in aDApTA, such as optimization, distribution, and materialization. Furthermore, due to difficulties in reuse third party tools, facing bugs and lack of documentation, only a couple were used.

6.3 Future works

Some of the complexity aspects of information management were addressed in this thesis. In future works, other aspects can be handled, and improvements can be considered in the aspects already treated. Furthermore, the approach, architecture and evaluations described can be evolve. Grouping the future works by topics, some of them are:

Feedback: the information from social media and the collaboration of the crowd can be used in aDApTA to evolve the integrated view of the situation to support decision making as discussed in Cordeiro et al. (2011b, 2011c, 2011d). In this perspective, unstructured data issues must be handled using approaches, such as those proposed in Moreira et al. (2013). In the same sense, data from sensors, human sensors, and even images from drones, are powerful sources to improve the dynamics of the situation awareness and need to be considered by aDApTA. Thus, new issues arise, such as noise variations and multi-sensor dynamics.

Uncertainty: this is a complexity aspect that has been addressed by the database community in the scope of dataspaces and uncertainty databases. In 1995, Motro discussed many issues that would be present in the future databases, such as imprecision, incompleteness, vagueness, inconsistency, and ambiguity. Later, Halevy et al. (2007) set

some principles discussing some solutions, such as keyword queries, ranked answers and data derivation through lineage. These issues are gaining importance with the increasing volume of data stored and available on the Web, most of them, generated by the Web of things, social, and media network. Addressing uncertainty aspects has become a requirement in information management where many issues are still open.

Heterogeneity: concerning the semantic heterogeneity level conflicts, there is a lack of research works evaluating the improvement on information integration using, so called, well-founded implemented domain ontology in LOD datasets. One of the main challenges is how to implement and apply the precision of the concepts modeled using a foundational ontology keeping the facilities of the Linked Data principles. The application of foundational ontologies for information integration and application in complex domains has already been discussed in Ferreira et al. (2010), Campos and Guizzardi (2010), Cordeiro et al. (2011e), Moreira et al. (2014), Moreira et al. (2015), and Ferreira et al. (2015). In this scenario, the semantic expressivity levels identified by aDApTA can be extended. Furthermore, to semantically enrich the incoming data, automated alignment approaches can be applied to create the mappings used in the *Annotated with Conceptual Description* activity.

Evaluation: in order to improve the evaluation of aDApTA extending its scope, a broader set of data sources of humanitarian logistics, such as those found in Sahana Eden[33], AidMatrix[34], Ushahidi[35], and iRevolutions[36] projects, can be used with a larger set of interlinking approaches. With this objective, some issues arise in the supporting architecture design, such as ETL job performance. Hence, it will be necessary to explore Big Data technologies and solutions to support the processing of provenance data in larger datasets. Other application domains can also be considered, such as command and control operations on conflict handling in a military area. Moreover, the classification of a larger set of interlinking tools can increase the heterogeneity levels of conflicts covered.

[33] http://eden.sahanafoundation.org

[34] http://www.aidmatrix.org

[35] http://www.ushahidi.com

[36] http://irevolution.net

References

AARNIO, P.; SEILONEN, I. **RDF Triple Stores as a Knowledge Management Technology for CBM Services**. In: IEEE International Conference on Emerging Technology & Factory Automation, Italy, 2013, **Proceedings**..., p. 10-13.

ACOSTA, M.; VIDAL, M.-E.; LAMPO, T.; CASTILLO, J.; RUCKHAUS, E. **ANAPSID: An adaptive query processing engine for SPARQL endpoints**. The Semantic Web – ISWC 2011. LNCS, Germany: Springer Berlin Heidelberg, v. 7031, p. 18-34.

AMARAL, G. C. M.; CAMPOS, M. L. M. **AQUAWARE: A Data Quality Support Environment for Data Warehousing**. In: Brazilian Symposium on Databases, Brazil, 2004, **Proceedings**..., p. 121-133.

ARAUJO, S.; HOUBEN, G.; SCHWABE, D.; HIDDERS, J. **Fusion – Visually Exploring and Eliciting Relationships in Linked Data**. The Semantic Web – ISWC 2010. LNCS, China: Springer Berlin Heidelberg, v. 6496, p. 1-15.

ARAUJO, S. F. C. **Data Integration over Distributed and Heterogeneous Data Endpoints.** 2014. 188 p. Ph.D. Thesis, Research School for Information and Knowledge Systems, Netherlands.

ASSAF, A.; LOUW, E.; SENART, A.; FOLLENFANT, C.; TRONCY, R.; TRASTOUR, D. **Improving Schema Matching with Linked Data**. In: First International Workshop on Open Data, France, 2012, **Proceedings**....

AXELROD, R.; COHEN, M. **Harnessing Complexity: Organizational Implications of a Scientific Frontier**. New York: Free Press, 2000, p. 208.

BARR, J. L.; BURTNER, E. R.; PIKE, W. A.; BOEK-PEDDICORD, A.; MINSK, B. S. **Gap Assessement in the Emergency Response Community**. **Department of Homeland Security.** EUA, 2010. Available at: <http://www.pnl.gov/main/publications/external/technical_reports/PNNL-19782.pdf>. Access in: 02/2015.

BARR, J. L.; BOEK-PEDDICORD, A. M.; BURTNER, E. R.; MAHY, H. A. **Current Domain Challenges in the Emergency Response Community**. In: International Conference at Information System for Crisis Response and Management, Portugal, 2011, **Proceedings**....

BARRAT, A.; BARTHÉLEMY, M.; VESPIGNANI, A. **Dynamical Processes on Complex Networks.** Cambridge University Press, 2008, p. 368.

BELLATRECHE, L.; PIERRA, G.; SARDET, E. **Evolution Management of Data Integration Systems by the Means of Ontological Continuity Principle**. Recent Trends in Information Reuse and Integration Book. Springer Vienna, 2012, p. 77-96.

BELLENGER, A. **Semantic Decision Support for Information Fusion Applications**. 2013. 222 p. Ph.D. Thesis, Institut National des Sciences Appliquées de Rouen, France.

BENBYA, H.; MCKELVEY, B. **Toward a Complexity Theory of Information Systems Development**. Information Technology & People, Emerald, USA, v. 19, n. 1, p. 12-34, 2006.

BENINI, A. A.; CONLEY, C. E.; SHDEED, R.; SPURWAY, K.; YARMOSHUK, M. **Integration of different data bodies for humanitarian decision support: an example from mine action.** Disasters, Wiley, v. 27, n. 4, p. 288-304, 2003.

BERNERS-LEE, T. **Putting the web back in Semantic Web** - ISWC 2005. Wrap-Up Session and RuleML Kickoff Sessions, Ireland. Available at: <http://www.w3.org/2005/Talks/1110-iswc-tbl>. Access in: 02/2015.

BERNERS-LEE, T. **Linked data**. W3C Design Issues. 2006. Available at: <http://www.w3.org/DesignIssues/LinkedData.html>. Access in: 02/2015.

BHAROSA, N.; JANSSEN, M. **Extracting Principles for Information Management Adaptability during Crisis Response: A Dynamic Capability View**. In: IEEE 43rd Hawaii International Conference on System Sciences, EUA, 2010, **Proceedings**..., p. 1-10.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. **Linked Data - The Story So Far**. International Journal on Semantic Web and Information Systems, IGI Global, v. 5, n. 3, p. 1-22, 2009.

BIZER, C.; SCHULTZ, A. **The R2R Framework Publishing and Discovering Mappings on the Web**. In: 1st International Workshop on Consuming Linked Data, China, 2010, **Proceedings**....

BLEIHOLDER, J.; NAUMANN, F. **Data fusion**. ACM Computing Surveys, v. 41, n. 1, p. 1-41, 2008.

BORGES, M. R. S.; BRÉZILLON, P.; PINO, J. A.; POMEROL, J.-CH. **Groupware System Design and the Context Concept.** Computer Supported Cooperative Work in Design I. LNCS, Springer Berlin Heidelberg. v. 3168, p. 45-54, 2005.

BORGES, M. R. S.; BRÉZILLON, P.; PINO, J. A.; POMEROL, J.-CH. **Dealing with the Effects of Context Mismatch in Group Work**. Decision Support Systems, v. 43, n. 4, p. 1692-1706, 2007.

BOUGUETTAYA, A.; BENATALLAH, B.; MEDJAHED, B.; OUZZANI, M.; HENDRA, L. **Adaptive web-based database communities**. USA: Information Modeling for Internet Applications, IGI Global, 2003, p. 277-298.

BRANDAO, S.; OLIVEIRA, J.; SOUZA, J. **Knowledge Representation with Autonomic Ontologies**. On the Move to Meaningful Internet Systems: OTM 2010 Workshops. 2010. LNCS, Greece: Springer Berlin Heidelberg, v. 6428, p. 635-644.

BRESLIN, J. G.; O'SULLIVAN, D.; PASSANT, A.; VASILIU, L. **Semantic Web Computing in Industry**. Computers in Industry, Emerald, v. 61, n. 8, p. 729-741, 2010.

BRÉZILLON, P.; POMEROL, J.-CH. **Contextual Knowledge Sharing and Cooperation in Intelligent Assistant Systems**. Le Travail Humain v. 62, n. 3, p. 223–246, 1999.

BROWN, D. E.; DUREN, B. G. **Conflicting Information Integration for Decision Support**. Decision Support Systems, Elsevier, v. 2, n. 4, p. 321-330, 1986.

BUNEMAN, P.; DAVIDSON, S. B. **Data provenance – the foundation of data quality**. 2010. Available at: <www.sei.cmu.edu/measurement/research/upload/Davidson.Pdf>. Access in: 02/2015.

CAMPOS, M. L. M.; GUIZZARDI, G. **GT-LinkedDataBR – Exposição, compartilhamento e conexão de recursos de dados abertos na Web (Linked Open Data).** Available at: <http://www.rnp.br/pd/gts2010-2011/gt_linkeddatabr.html>. Access in: 09/2013.

CORDEIRO, K. F.; FARIA, F. F.; PEREIRA, B. O.; FREITAS, A.; RIBEIRO, C. E.; FREITAS, J. V. V. B.; BRINGUENTE, A. C.; ARANTES, L. O.; CALHAU, R.; ZAMBORLINI, V.; CAMPOS, M. L. M.; GUIZZARDI, G. **An approach for managing and semantically enriching the publication of Linked Open Governmental Data**. In: 3rd Workshop in Applied Computing for Electronic Government, Brazilian Symposium on Database, Brazil, 2011a, **Proceedings**..., p. 82-95.

CORDEIRO, K. F.; MARINO, T.; CAMPOS, M. L. M.; BORGES, M. R. S. **Use of Linked Data in the Design of Information Infrastructure for Collaborative Emergency Management System**. In: IEEE 15th International Conference on Computer Supported Cooperative Work in Design, Switzerland, 2011b, **Proceedings**..., p. 764-771.

CORDEIRO, K. F.; CAMPOS, M. L. M.; BORGES, M. R. S. **Empowering Citizens and Government with Collaboration on Linked Open Data**. In: Workshop on Semantics in Governance and Policy Modelling at Extended Semantic Web Conference, Greece, 2011c, **Proceedings**....

CORDEIRO, K. F.; CAMPOS, M. L. M.; BORGES, M. R. S. **Collaboration Issues on Linking Open Data**. In: Poster Session at Extended Semantic Web Conference, Greece, 2011d, **Proceedings**....

CORDEIRO, K. F.; ENGELBRECHT, A.; OLIVEIRA, J.; CAMPOS, M. L. M. **Cooperative Learning of Advanced Conceptual Modeling Principles using Semantic Wiki**. IEEE Multidisciplinary Engineering Education Magazine, v. 6, n. 3, p. 19-30, 2011e.

CORDEIRO, K. F.; CAMPOS, M. L. M.; BORGES, M. R. S. **Adaptive Integration of Information Supporting Decision Making: A Case on Humanitarian Logistic**. In: 11th International Conference on Information Systems for Crisis Response and Management, USA, 2014a, **Proceedings**..., p. 225-229.

CORDEIRO, K. F.; CAMPOS, M. L .M.; BORGES, M. R. S. **Adaptive Approach for Information Integration supported by Linked Open Data**. In: Poster Session of Early Career Seminar at 8th International Conference on Formal Ontology in Information Systems, Brazil, 2014b, **Proceedings**....

CORDEIRO, K. F.; CAMPOS, M. L. M.; BORGES, M. R. S. **aDApTA - Adaptive Approach to Information Integration in Dynamic Environments**. Journal of Computers in Industry, Accepted for publication, 2015.

CRUZ, S. M. S.; CAMPOS, M. L. M.; MATTOSO, M. **Towards a Taxonomy of Provenance in Scientific Workflow Management Systems**. In: IEEE Congress on Services, USA, 2009, **Proceedings**..., p. 259-266.

DE LA CERDA, J.; CAVALCANTI, M. C. **Registro de procedência de ligações RDF em Dados Ligados**. In: Joint V Seminar on Ontology Research in Brazil and VII International Workshop on Metamodels, Ontologies and Semantic Technologies, Brazil, 2012, **Proceedings**....

DE VETTOR, P.; MRISSA, M.; PEDRINACI, C. **Context mediation as a linked service**. Service-Oriented and Cloud Computing. 2012. LNCS, Italy: Springer Berlin Heidelberg, v. 7592, p. 210-211.

DE VETTOR, P.; MRISSA, M.; BENSLIMANE, D.; BERBAR, S. **A Service Oriented Architecture for Linked Data Integration**. In: IEEE 8th International Symposium on Service-Oriented Systems, United Kingdom, 2014, **Proceedings**..., p. 198-203.

Department of Social Welfare & Development of Philippines. **Humanitarian Response**. 2014. Available at: <http://www.humanitarianresponse.info/operations/philippines/dataset>. Access in: 02/2015.

DIMOU, A.; MIEL, V.S.; PIETER, C.; RUBEN, V.; ERIK, M.; DE WALLE, R. V. **RML: a Generic Language for Integrated RDF Mappings of Heterogeneous Data**. In: 7th Workshop on Linked Data on the Web, Korea, 2014, **Proceedings**....

DINIZ, V. B.; BORGES, M. R. S.; GOMES, J. O.; CANÓS, J. H. **Knowledge management support for collaborative emergency response**. In: IEEE 9th International Conference on Compter Supported Cooperative Work in Design, United Kingdom, 2005, **Proceedings**..., p. 1188-1193.

DIVIDINO, R.; SIZOV, S.; STAAB, S.; SCHUELER, B. **Querying for Provenance, Trust, Uncertainty and Other Meta Knowledge in RDF**. Journal of Web Semantics, Elsevier, v. 7, n. 3, p. 204-219, 2009.

FAHLAND, D.; GLÄßER, T. M.; QUILITZ, B.; WEIßLEDER, S.; LESER, U. **HUODINI – Flexible Information Integration for Disaster Management.** In: International Conference at Information System for Crisis Response and Management, China, 2007, **Proceedings**…, p 1-8.

FERRARA, A.; NIKOLOV, A.; SCHARFFE, F. **Data Linking for the Semantic Web**. International Journal on Semantic Web and Information Systems, IGI Global, v. 7, n. 3, p. 46-76, 2011.

FERREIRA, M. I. G. B.; CORDEIRO, K. F.; OLIVEIRA, J.; CAMPOS, M. L. M. **OntoEmerge: Construção de uma Ontologia Core para a Área de Emergências Baseada em Ontologia de Fundamentação**. In: Poster Session at III Seminar on Ontology Research in Brazil, Brazil, 2010, **Proceedings**....

FERREIRA, M. I. G. B.; BRAGA, B. F. B.; SALES, T. P.; CORDEIRO, K. F.; CAMPOS, M. L. M.; MOREIRA, J. L. R.; BORGES, M. R. S. **OntoEmergePlan: variability of emergency plans supported by a domain ontology**. In: 12th International Conference on Information Systems for Crisis Response and Management, Norway, 2015, **Proceedings**….

FRANKLIN, M.; HALEVY, A. **From databases to dataspaces: a new abstraction for information management**. ACM Sigmod Record, p. 1-7, 2005.

FREITAS, A.; KAMPGEN, B.; OLIVEIRA, J. G.; O'RIAIN, S.; CURRY, E. **Representing Interoperable Provenance Descriptions for ETL Workflows.** In: 3rd International Workshop on Role of Semantic Web in Provenance Management, Crete, 2012, **Proceedings**….

GAO, W. **An Approach to Formalizing Ontology Driven Semantic Integration: Concepts, Dimensions and Framework**. 2012. 169 p. Ph.D. Thesis, The Florida State University, USA.

GIANNOTTI, M. A. **Ontologies development for humanitarian logistics support systems based on geographic information services: a food bank application (Portuguese)**. 2011. 197 p. Doctoral Thesis, Universidade de São Paulo, Brazil.

GRAUBE, M.; PFEFFER, J.; ZIEGLER, J.; URBAS, L. **Linked Data as Integrating Technology for Industrial Data**. In: IEEE 14th International Conference in Network-Based Information Systems, Albania, 2011, **Proceedings**…, p. 162-167.

GROTH, P.; LOIZOU, A.; GRAY, A.; GOBLE, C.; HARLAND, L.; PETTIFER, S. **API-centric Linked Data Integration: The Open PHACTS Discovery Platform Case Study**. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, In Press, Elsevier, v. 29, 2014.

HADDOW, G.; BULLOCK, J.; COPPOLA, D. **Introduction to Emergency Management**. Butterworth-Heinemann, 2011, p. 424.

HALEVY, A. Y. **Theory of Answering Queries Using Views**. SIGMOD Record, v. 29, n. 4, p. 40-47, 2000.

HALEVY, A; FRANKLIN, M; MAIER, D. **Principles of Dataspace Systems**. In: 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, USA, 2006, **Proceedings**....

HARISPE, S.; RANWEZ, S.; JANAQI, S.; MONTMAIN, J. **Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis**. CoRR, v. 1310.1285, 2013a.

HARISPE, S.; RANWEZ, S.; JANAQI, S.; MONTMAIN, J. **Semantic Measures Based on RDF Projections: Application to Content-Based Recommendation Systems**. On the Move to Meaningful Internet Systems: OTM Conferences. 2013b. LNCS, Austria: Springer Berlin Heidelberg, v. 8185, p. 606-615.

HARTIG, O. **Provenance Information in the Web of Data**. In: Linked Data on the Web at International World Wide Web Conference, Spain, 2009, **Proceedings**....

HEATH, T.; BIZER, C. **Linked Data: Evolving the Web into a Global Data Space**. Synthesis Lectures on the Semantic Web: Theory and Technology, v. 1, n. 1, p. 1-136, 2011.

HEDELER, C.; BELHAJJAME, K.; MAO, L.; GUO, C. **DSToolkit: An Architecture for Flexible Dataspace Management**. Transactions on Large-Scale Data- and Knowledge-Centered Systems, v. 5, p. 126-157, 2012.

HELLINGRATH, B.; WIDERA, A. **Survey on Major Challenges in Humanitarian Logistics**. In: International Conference at Information System for Crisis Response and Management, Portugal, 2011, **Proceedings….**

HRISTIDIS, V.; SHU-CHING, Chen; TAO, Li; LUIS, S.; YI, Deng. **Survey of data management and analysis in disaster situations**. Journal of Systems and Software, v. 83, n. 10, p. 1701-1714, 2010.

ISELE, R.; JENTZSCH, A.; BIZER, C. **Silk server - adding missing links while consuming linked data**. In: 1st International Workshop on Consuming Linked Data, China, 2010, **Proceedings**....

JAIN, P.; HITZLER, P.; SHETH, A. P.; VERMA, K.; YEH, P. Z. **Ontology Alignment for Linked Open Data**. In: 9th International Semantic Web Conference, China, 2010, **Proceedings**..., v. 6496, p. 402-417.

JAIN, P.; YEH, P.Z.; VERMA, K.; VASQUEZ, R. G.; DAMOVA, M.; HITZLER, P.; SHETH. A. P. **Contextual ontology alignment of LOD with an upper ontology: a case study with proton**. In: 8th Extended Semantic Web Conference on the Semantic Web: Research and Applications, Greece, 2011, **Proceedings**…, p. 80-92.

JIACAI, Ni; GUOLIANG, Li; JUN, Zhang; LEI, Li; JIANHUA, Feng. **Adapt: adaptive database schema design for multi-tenant applications**. In: 21st ACM International Conference on Information and Knowledge Management, USA, 2012, **Proceedings**…, p. 2199-2203.

JUN, Yang; QING, Li; YUETING, Zhuang. **A Self-Adaptive Semantic Schema Mechanism for Multimedia Databases**. In: SPIE Electronic Imaging and Multimedia Technology III, China, 2002, **Proceedings**….

KHAN, H.; VASILESCU, L.; KHAN, A. **Disaster Management CYCLE – A Theoretical Approach**. Journal Management & Marketing, v. 6, p. 43-50, 2008.

KHAZANKIN, R.; DUSTDAR, S. **On Adaptive Integration of Web Data Sources into Applications**. In: International Workshop Innovation Information Technologies: Theory and Practice, Germany, 2010 , **Proceedings**….

KIMBALL, R.; MOSS, M. **The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling**. Wiley, 2013, p. 600.

KNAP, T.; MICHELFEIT, J.; DANIEL, J.; JERMAN, P.; RYCHNOVSKÝ, D.; SOUKUP, T.; NEČASKÝ, M. **ODCleanStore: A Framework for Managing and Providing Integrated Linked Data on the Web**. In: Web Information Systems Engineering, 2012, **Proceedings**…, p. 815-816.

KONDYLAKIS, H.; FLOURIS, G.; PLEXOUSAKIS, D. **Ontology and Schema Evolution in Data Integration: Review and Assessment**. In: On the Move Federated Conferences & Workshops, Portugal, 2009, **Proceedings**…, p. 932-947.

LANGEGGER, A.; WÖß, W.; BLÖCHL, M. **A Semantic Web Middleware for Virtual Data Integration on the Web**. The Semantic Web: Research and Applications. 2008. LNCS, Spain: Springer Berlin Heidelberg, v. 5021, p. 493-507.

LAUDY, C.; GANASCIA, J-G.; SEDOGBO, C. **High-level Fusion based on Conceptual Graphs**. In: 10th International Conference on Information Fusion, Canada, 2007, **Proceedings**…, p. 1-8.

LEHMANN, J.; ISELE, R.; JAKOB, M.; JENTZSCH, A.; KONTOKOSTAS, D.; MENDES, P. N.; HELLMANN, S.; MORSEY, M.; VAN KLEEF, P.; AUER, S.; BIZER, C. **DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia**. Semantic Web Journal, Under Review, 2013.

LENZERINI, M. **Data integration: A theoretical perspective**. In: Symposium on Principles of Database Systems, USA, 2002, **Proceedings**…, p. 233-246.

LE-PHUOC, D.; DAO-TRAN, M.; PARREIRA, J. X.; HAUSWIRTH, M. **A native and adaptive approach for unified processing of linked streams and linked data**. In: 10th International Semantic Web Conference, Germany, 2011, **Proceedings**…, p. 23-27.

LINDELL, M. K.; PRATER, C.; PERRY, R. W. **Introduction to Emergency Management**. John Wiley & Sons, 2006, p. 624.

LÓSCIO, B. F. **Managing the Evolution of XML-based Mediation Queries**. 2003. 159 p. Doctoral Thesis, Universidade Federal de Pernambuco, Brazil.

LYNDEN, S.; ISAO, Kojima; AKIYOSHI, Matono; YUSUKE, Tanimura. **Adaptive integration of distributed semantic web data**. Databases in Networked Information Systems. 2010. LNCS, Japan: Springer Berlin Heidelberg, v. 5999, p. 174-193.

MADHAVAN, J.; JEFFERY, S.; COHEN, S.; DONG, X. **Web-scale data integration: You can only afford to pay as you go**. In: 3rd Biennial Conference on Innovative Data Systems Research, USA, 2007, **Proceedings**..., p. 342-350.

MAHFOUDH, M.; FORESTIER, G.; THIRY, L.; HASSENFORDER, M. **Consistent ontologies evolution using graph grammars**. Knowledge Science, Engineering and Management. 2013. LNCS, China: Springer Berlin Heidelberg, v. 8041, p. 64-75.

MARJIT, U.; SHARMA, K.; BISWAS, U. **Provenance Representation and Storage Techniques in Linked Data: A State-of-the-art Survey**. International Journal of Computer Applications, v. 38, n. 9, p. 23-28, 2012.

MENDES, P. N.; MUHLEISEN, H.; BIZER, C. **Sieve: Linked Data Quality Assessment and Fusion**. In: 2nd International Workshop on Linked Web Data Management at EDBT/ICDT Joint Conference, ACM, Germany, 2012, **Proceedings**....

MENDONÇA, R. R. **Uma abordagem para coleta e publicação de dados de proveniência no contexto de Linked Data.** 2013. 145 p. Master Dissertation, Universidade Federal do Rio de Janeiro, Brazil.

MENDONÇA, R. R.; CRUZ, S. M. S.; DE LA CERDA, J. F. S. M.; CAVALCANTI, M. C.; CORDEIRO, K. F.; CAMPOS, M. L. M. **LOP – Capturing and Linking Open Provenance on LOD Cycle**. In: 5th International Workshop on Semantic Web Information Management, EUA, 2013, **Proceedings**.... SIGMOD.

MICHELFEIT, J. **Linked Data Integration**. 2013. 93 p. Master Dissertation, Charles University in Prague, Czech Republic.

MICHELFEIT, J.; KNAP, T.; NEČASKÝ, M. **Linked Data Integration with Conflicts.** Journal of Web Semantic, Preprint, 2014.

MIJOVIC, V.; JANEV, V.; VRANE, S. **Main Challenges in Using LOD in Emergency Management**. In: International Workshop on Database and Expert Systems Applications, USA, 2013, **Proceedings**..., p. 21-25.

MITCHELL, M. **Complexity: A Guided Tour**. USA: Oxford University Press, 2009, p. 368.

MOREAU, L.; CLIFFORD, B.; FREIRE, J.; FUTRELLE, J.; GIL, Y.; GROTH, P.; KWASNIKOWSKA, N.; MILES, S.; MILES, P.; MYERS, J.; PLALE, B.; SIMMHAN, Y.; STEPHAN, E.; BUSSCHE, J. V. D. **The open provenance model core specification**. Future Generation Computer Systems, Elsevier, v. 27, n. 6, p. 743-756, 2011.

MOREIRA, J. L.; CORDEIRO, K. F.; CAMPOS, M. L. M. **JointOLAP - Sistema de Informação para Exploração Conjunta de Dados Estruturados e Textuais: Um estudo de caso no setor elétrico.** In: Simpósio Brasileiro de Sistemas de Informação, Brazil, 2013, **Proceedings**….

MOREIRA, J. L.; CORDEIRO, K. F.; CAMPOS, M. L. M. **OntoWarehousing – Multidimensional Design Supported by a Foundational Ontology: A Temporal Perspective**. In: 16th International Conference on Data Warehousing and Knowledge Discovery, Germany, 2014, **Proceedings**….

MOREIRA, J. L.; CORDEIRO, K. F.; CAMPOS, M. L. M.; BORGES, M. R. S. **Hybrid Multidimensional Design for Heterogeneous Data Supported by Ontological Analysis: An Application Case in the Brazilian Electric System Operation.** In: 4th Workshop on Energy Data Management at EDBT/ICDT Joint Conference, Belgium, 2015, **Proceedings**….

MORI, M.; CLEVE, A. **Feature-based Adaptation of Database Schemas**. Model-Based Methodologies for Pervasive and Embedded Software. 2013. LNCS, Germany: Springer Berlin Heidelberg, v. 7706, p. 85-105.

MORIN, E. **Introdução ao Pensamento Complexo**. Brazil: Sulina, 2007, p. 120.

MOTRO, A. **Management of Uncertainty in Database Systems**. Modern Database Systems. USA: ACM Press/Addison-Wesley, p. 457-476.

MRISSA, M.; SELLAMI, M.; DE VETTOR, P.; BENSLIMANE, D.; DEFUDE, B. **A Decentralized Mediation-as-a-Service Architecture for Service Composition**. In: IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, Tunisia, 2013, **Proceedings**…, p. 80-85.

NEČASKÝ, M.; KLÍMEK, J.; MYNARZ, J.; KNAP, T.; SVÁTEK, V.; STÁRKA, J. **Linked data support for filing public contracts**. Computers in Industry, Elsevier, v. 65, n. 5, p. 862-877, 2014.

NICOLLE, C.; CRUZ, C. **Adaptive Integration of Information**. In: 5th International Conference on Web Information Systems and Technologies, Portugal, 2009, **Proceedings**…, p. 115-118.

NWANA, H. S.; NDUMU, D. T. **An Introduction to Agent Technology**. Software Agents and Soft Computing Towards Enhancing Machine Intelligence. 1997. LNCS, United Kingdom: Springer-Verlag, v. 1198, p. 1-26.

OMG (Object Management Group), **Business Process Model and Notation (BPMN)**, 2011. Available at: <http://www.omg.org/spec/BPMN/2.0>. Access in: 02/2015.

OXLEY, M.E.; THORSEN, S. N. **Fusion and integration. What's the difference?** In: 7th International Conference on Information Fusion, 2004, **Proceedings**…, v. 1, p. 429-434.

PASSANT, A. **Semantic Web Technologies for Enterprise 2.0**. Studies on the Semantic Web, IOS Press, 2010, v. 9, p. 348.

RAM, S.; PARK, J. **Semantic Conflict Resolution Ontology (SCROL): An Ontology for Detecting and Resolving Data and Schema-Level Semantic Conflicts**. In: IEEE Transactions on Knowledge and Data Engineering, 2004, **Proceedings**..., v. 16, n. 2, p. 189-202.

ROBINSON, I.; WEBBER, J.; EIFREM, E. **Graph Databases**. O`Reilly Media, 2013, p. 224.

RODDICK, G. F. **A survey of schema versioning issues for database systems**. Information and Software Technology, v. 37, n. 7, p. 383-393, 1995.

RODRIGUEZ, M. A.; NEUBAUER, P. **Constructions from Dots and Lines**. Bulletin of the American Society for Information Science and Technology, Wiley, v. 36, n. 6, p. 35-41, 2010.

SAHOO, S. S.; SHETH, A.; HENSON, C. **Semantic Provenance for eScience: Managing the Deluge of Scientific Data**. In: IEEE Internet Computing, 2008, **Proceedings**..., v. 4, n. 12, p. 46-54.

SALGADO, A. C.; LÓSCIO, B.; BATISTA, M. C. M.; BELIAN, R. B.; PIRES, C. E.; SOUZA, D. Y. **The Data Integration Research Group at UFPE**. Journal of Information and Data Management, v. 2, p. 109-122, 2011.

SALLES, M. A. V.; DITTRICH, J.-P.; KARAKASHIAN, S. K.; GIRARD, O. R.; BLUNSCHI, L. **ITRAILS: Pay-as-you-go Information Integration in Dataspaces**. In: International Conference on Very Large Data Bases, Austria, 2007, **Proceedings**..., p. 663-674.

SALTOR, F.; CASTELLANOS, M.; GARCÍA-SOLACO, M. **Suitability of datamodels as canonical models for federated databases**. ACM SIGMOD Record, v. 20, n. 4, p. 44-48, 1991.

SANTOS, R. S.; BORGES, M. R. S.; CANÓS, J. H.; GOMES, J. O. **The Assessment of Information Technology Maturity in Emergency Response Organizations**, Group Decision and Negotiation, v. 20, n. 5, p. 593-613, 2011.

SCHARFFE, F.; EUZENAT, J. **MeLinDa: an interlinking framework for the web of data**. Computing Research Repository. CoRR, n. 1107.4502, 2011.

SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H. **State of the LOD Cloud 2014.** Availabe at: <http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/>. Access in: 03/2015.

SCHONENBERG, H.; MANS, R.; RUSSELL, N.; MULYAR, N.; VAN DER AALST, W.; **Process Flexibility: A Survey of Contemporary Approaches**. Advances in Enterprise Engineering I. Lecture Notes in Business Information Processing. Springer Berlin Heidelberg, v. 10, p. 16-30, 2008.

SCHULTZ, A.; MATTEINI, A.; ISELE, R.; BIZER, C.; BECKER, C. **LDIF: Linked data integration framework**. In: 2nd International Workshop on Consuming Linked Data, Germany, 2011, **Proceedings**….

SCHULZ, A.; PAULHEIM, H.; PROBST, F. **Crisis Information Management in the Web 3.0 Age**. In: International Conference on Information Systems for Crisis Response and Management, Canada, 2012a, **Proceedings**…, p. 1-5.

SCHULZ, A.; DOWELING, S.; PROBST, F. **Integrating Process Modeling and Linked Open Data to Improve Decision Making in Disaster Management**. In: ACM Conference on Computer Supported Cooperative Work, 2012b, **Proceedings**…, p. 12-15.

SELLAMI, M.; DE VETTOR, P.; MRISSA, M.; BENSLIMANE, D.; DEFUDE, B. **DMaaS : Syntactic, Structural and Semantic Mediation for Service Composition**. In: International Journal of Autonomous and Adaptive Communications Systems, Inderscience Publishers, 2014, **Proceedings**….

SIAU, K.; TIAN, Y.**Supply chain integration: architecture and enabling technologies**. The Journal of Computer Information Systems, v. 44, p. 67-72, 2004.

SILVA, V. S.; CAMPOS, M. L. M.; SILVA, J. C. P.; CAVALCANTI, M. C. **An Approach for the Alignment of Biomedical Ontologies Based on Foundational Ontologies**. Journal of Information and Data Management, v. 2, p. 557-572, 2011.

SIMON, H. A. **The Architecture of Complexity**. In: The American Philosophical Society, 1962, **Proceedings**....

SNOWDEN, D. J.; BOONE, M. E. **A Leaders Framework for Decision Making - Wise Executive Tailor their Approach to Fit the Complexity of the Circumstances They Face**. Harvard Business Review, v. 85, p. 68, 2007.

TAHERIYAN, M.; KNOBLOCK, C. A.; SZEKELY, P.; AMBITE, J. L. **A Graph based Approach to Learn Semantic Descriptions of Data Sources**. The Semantic Web – ISWC 2013. LNCS, Australia: Springer Berlin Heidelberg, v. 8218, p. 607-623.

THOLLOT, R. **Dynamic Situation Monitoring and Context-Aware BI Recommendations**. 2012. 172 p. Ph.D. Thesis, Ecole Centrale Paris, France.

United Nations Development Program in Phillipines. **Typhoon Haiyan (Yolanda) Strategic Response Plan.** Report. 2013. Available at: <https://docs.unocha.org/sites/dms/CAP/SRP_2013-2014_Philippines_Typhoon_Haiyan.pdf>. Access in: 02/2015.

United Nations Office for the Coordination of Humanitarian Affairs. **Philippines: Typhoon Haiyan Action Plan.** Report. 2013. Available at: <http://www.unocha.org/cap/appeals/philippines-typhoon-haiyan-action-plan-november-2013>. Access in: 02/2015.

United Nations Office for the Coordination of Humanitarian Affairs. 2014. **HumanitarianResponse.info**. Available at: <http://www.humanitarianresponse.info/operations/philippines/dataset>. Access in: 11/2014.

VOLZ, J.; BIZER, C.; GAEDKE, M.; KOBILAROV, G. **Silk - A Link Discovery Framework for the Web of Data**. In: Workshop about Linked Data on the Web, Spain, 2009, **Proceedings**….

WIDERA, A.; DIETRICH, H. -A.; HELLINGRATH, B.; BECKER, J. **Understanding Humanitarian Supply Chains – Developing an Integrated Process Analysis Toolkit**. In: International Conference on Information Systems for Crisis Response and Management, Germany, 2013, **Proceedings**…, p. 210-219.

WÖLGER, S.; SIORPAES, K.; BÜRGER, T.; SIMPERL, E.; THALER, S.; HOFER, C. **A Survey on Data Interlinking Methods**. Technical Report. Semantic Technology Institute. University of Innsbruck, Austria, 2011.

YONGTAO, Ma.; TRAN, T. **TYPiMatch: Type-specific Unsupervised Learning of Keys and Key Values for Heterogeneous Web Data Integration**. In: 6th ACM International Conference on Web Search and data Mining, Italy, 2013, **Proceedings**…, p. 325-334.

YONGTAO, Ma.; TRAN, T.; BICER, V. **TYPifier: Inferring the Type Semantics of Structured Data**. In: International Conference on Data Engineering, Australia, 2013, **Proceedings**…, p. 201-217.

YONGTAO, Ma. **Effective Instance Matching for Heterogeneous Structured Data**. 2014. 161 p. Master Dissertation, Karlsruhe Institute of Technology, Germany.

ZHAO, J.; HARTIG, O. **Towards Interoperable Provenance Publication on the Linked Data Web**. In: 5th Linked Data on the Web Workshop, France, 2012, **Proceedings**….

Appendices

APPENDIX A – HUMANITARIAN LOGISTIC ONTOLOGY

In order to provide a better visualization of the Humanitarian Logistic Ontology developed for the prototype used in this thesis, the OWL code and diagram are listed below. Also, the OWL code and the Ontology file are available at aDApTA web page (http://greco.ppgi.ufrj.br/lodbr/index.php/principal/adapta). The ontology was built using the Enterprise Architecture Software[37].

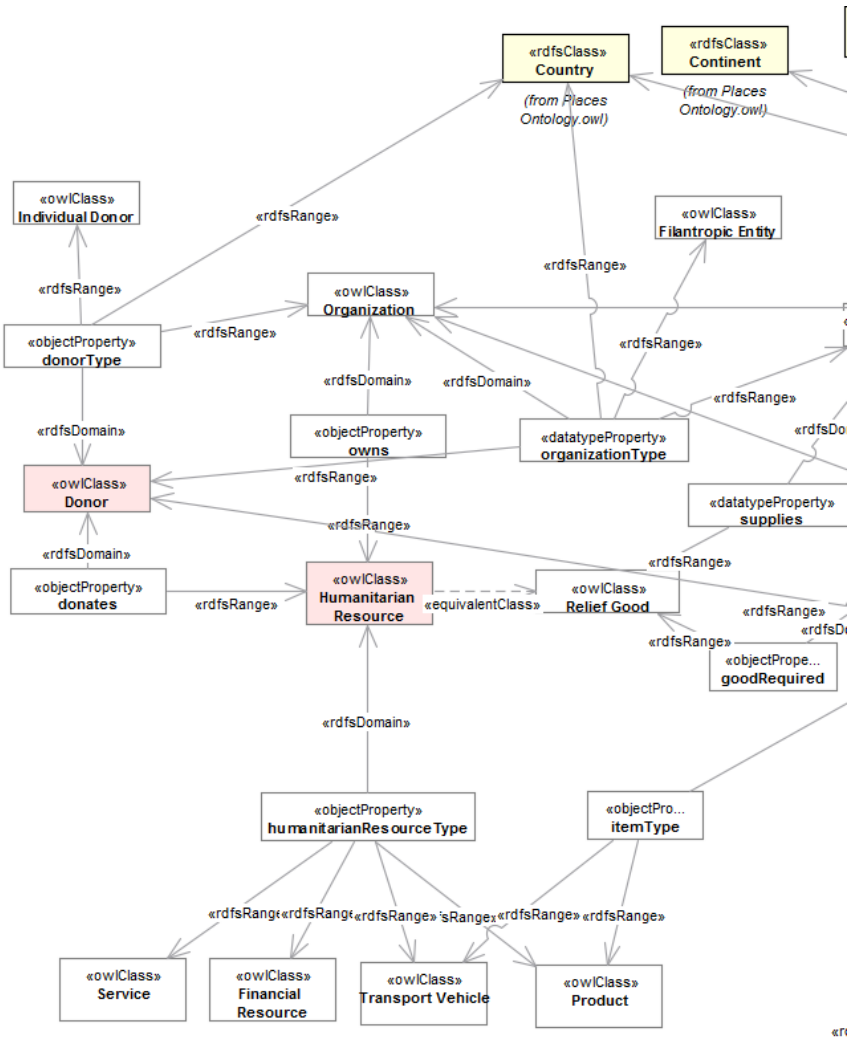


Figure 75 - Humanitarian Logistic Ontology Diagram - Part I

[37] http://www.sparxsystems.com.au/

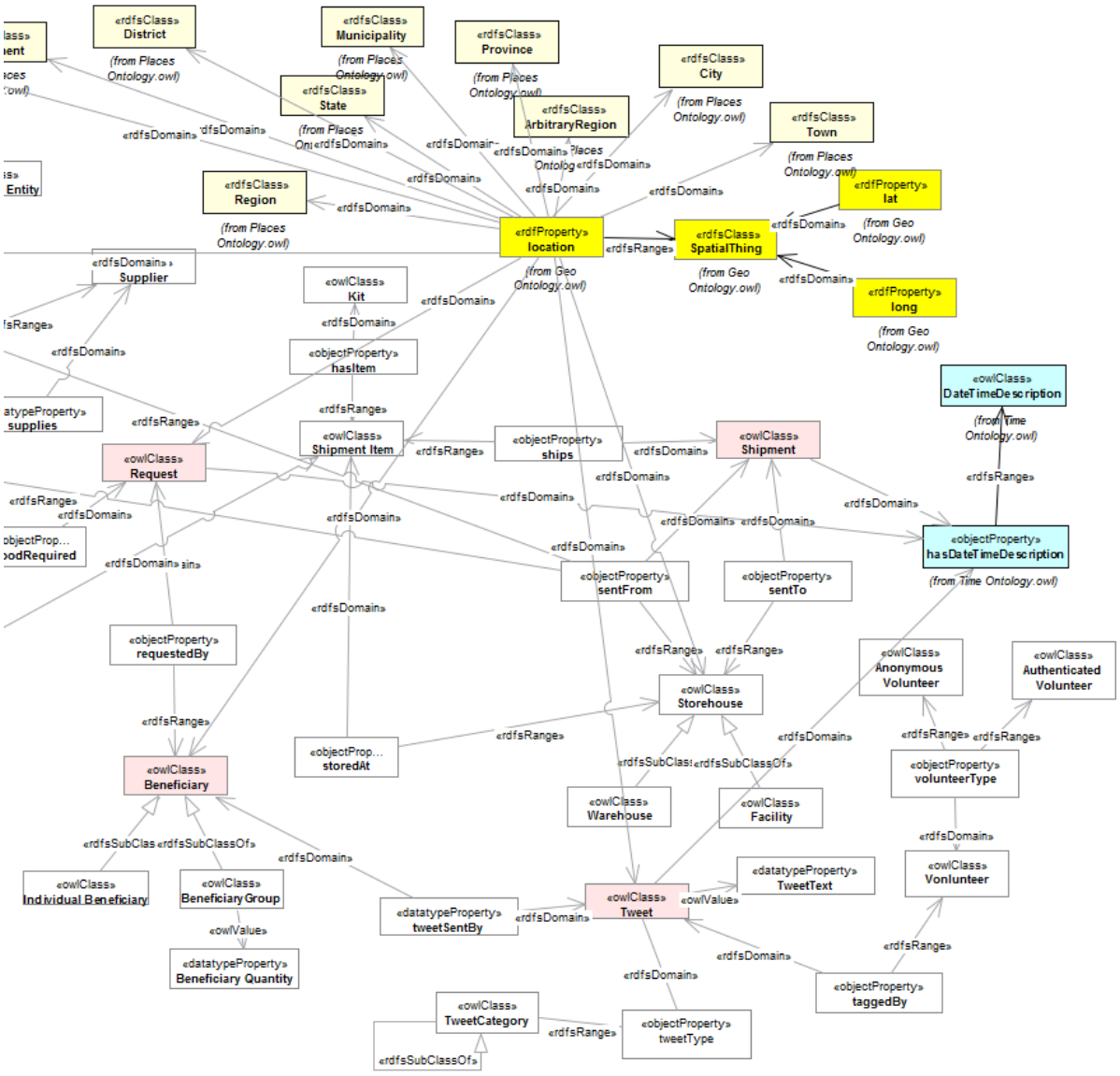


Figure 76: Humanitarian Logistic Ontology Diagram - Part II

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#">
  <owl:Class rdf:about="#AnonymousVolunteer" rdf:ID="Anonymous Volunteer" />
  <owl:Class rdf:about="#AuthenticatedVolunteer" rdf:ID="Authenticated Volunteer" />
  <owl:Class rdf:about="#Beneficiary" rdf:ID="Beneficiary" />
  <owl:Class rdf:about="#BeneficiaryGroup" rdf:ID="BeneficiaryGroup">
    <rdfs:subClassOf rdf:resource="#Beneficiary" />
  </owl:Class>
  <owl:Class rdf:about="#Donor" rdf:ID="Donor" />
  <owl:Class rdf:about="#Facility" rdf:ID="Facility">
    <rdfs:subClassOf rdf:resource="#Storehouse" />
  </owl:Class>
  <owl:Class rdf:about="#FilantropicEntity" rdf:ID="Filantropic Entity" />
  <owl:Class rdf:about="#FinancialResource" rdf:ID="Financial Resource" />
  <owl:Class rdf:about="#HumanitarianResource" rdf:ID="Humanitarian Resource">
    <owl:equivalentClass rdf:resource="#Good" />
  </owl:Class>
  <owl:Class rdf:about="#IndividualBeneficiary" rdf:ID="Individual Beneficiary">
```

```
    <rdfs:subClassOf rdf:resource="#Beneficiary" />
  </owl:Class>
  <owl:Class rdf:about="#IndividualDonor" rdf:ID="Individual Donor" />
  <owl:Class rdf:ID="Infrastructure Damage" />
  <owl:Class rdf:about="#Kit" rdf:ID="Kit" />
  <owl:Class rdf:ID="Need" />
  <owl:Class rdf:about="#Organization" rdf:ID="Organization" />
  <owl:Class rdf:about="#Product" rdf:ID="Product" />
  <owl:Class rdf:about="#Good" rdf:ID="Relief Good" />
  <owl:Class rdf:about="#Request" rdf:ID="Request" />
  <owl:Class rdf:ID="Request for Help">
    <owl:equivalentClass rdf:resource="#Request" />
  </owl:Class>
  <owl:Class rdf:about="#Service" rdf:ID="Service" />
  <owl:Class rdf:about="#Shipment" rdf:ID="Shipment" />
  <owl:Class rdf:about="#ShipmentItem" rdf:ID="Shipment Item" />
  <owl:Class rdf:about="#Storehouse" rdf:ID="Storehouse" />
  <owl:Class rdf:about="#Supplier" rdf:ID="Supplier" />
  <owl:Class rdf:about="#Transport" rdf:ID="Transport Vehicle" />
  <owl:Class rdf:about="#Tweet" rdf:ID="Tweet" />
  <owl:Class rdf:about="#TweetCategory" rdf:ID="TweetCategory">
    <rdfs:subClassOf rdf:resource="#TweetCategory" />
  </owl:Class>
  <owl:Class rdf:about="#TweetSubCategory" rdf:ID="TweetSubCategory" />
  <owl:Class rdf:about="#Vonlunteer" rdf:ID="Vonlunteer" />
  <owl:Class rdf:about="#Warehouse" rdf:ID="Warehouse">
    <rdfs:subClassOf rdf:resource="#Storehouse" />
  </owl:Class>
  <owl:DatatypeProperty rdf:about="#QuantityBeneficiaries" rdf:ID="Beneficiary
Quantity" />
  <owl:ObjectProperty rdf:ID="Request for Help Type" />
  <owl:ObjectProperty rdf:ID="TweetCategoryType" />
  <owl:DatatypeProperty rdf:about="#TweetText" rdf:ID="TweetText" />
  <owl:ObjectProperty rdf:ID="donates">
    <rdfs:domain rdf:resource="#Donor" />
    <rdfs:range rdf:resource="#HumanitarianResource" />
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="donorType">
    <rdfs:range rdf:resource="#Organization" />
    <rdfs:range rdf:resource="#IndividualDonor" />
    <rdfs:domain rdf:resource="#Donor" />
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="goodRequired">
    <rdfs:domain rdf:resource="#Request" />
    <rdfs:range rdf:resource="#Good" />
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="hasItem">
    <rdfs:range rdf:resource="#ShipmentItem" />
    <rdfs:domain rdf:resource="#Kit" />
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="humanitarianResourceType">
    <rdfs:range rdf:resource="#FinancialResource" />
    <rdfs:domain rdf:resource="#HumanitarianResource" />
    <rdfs:range rdf:resource="#Product" />
    <rdfs:range rdf:resource="#Transport" />
    <rdfs:range rdf:resource="#Service" />
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="itemType">
    <rdfs:range rdf:resource="#Product" />
    <rdfs:domain rdf:resource="#ShipmentItem" />
    <rdfs:range rdf:resource="#Transport" />
  </owl:ObjectProperty>
  <owl:DatatypeProperty rdf:ID="organizationType">
    <rdfs:range rdf:resource="#Donor" />
    <rdfs:range rdf:resource="#FilantropicEntity" />
```

```
    <rdfs:range rdf:resource="#Supplier" />
    <rdfs:domain rdf:resource="#Organization" />
  </owl:DatatypeProperty>
  <owl:ObjectProperty rdf:ID="owns">
    <rdfs:domain rdf:resource="#Organization" />
    <rdfs:range rdf:resource="#HumanitarianResource" />
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="requestedBy">
    <rdfs:domain rdf:resource="#Request" />
    <rdfs:range rdf:resource="#Beneficiary" />
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="sentFrom">
    <rdfs:domain rdf:resource="#Shipment" />
    <rdfs:range rdf:resource="#Storehouse" />
    <rdfs:range rdf:resource="#Donor" />
    <rdfs:range rdf:resource="#Organization" />
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="sentTo">
    <rdfs:range rdf:resource="#Storehouse" />
    <rdfs:domain rdf:resource="#Shipment" />
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="ships">
    <rdfs:domain rdf:resource="#Shipment" />
    <rdfs:range rdf:resource="#ShipmentItem" />
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="storedAt">
    <rdfs:range rdf:resource="#Storehouse" />
    <rdfs:domain rdf:resource="#ShipmentItem" />
  </owl:ObjectProperty>
  <owl:DatatypeProperty rdf:ID="supplies">
    <rdfs:range rdf:resource="#Good" />
    <rdfs:domain rdf:resource="#Supplier" />
  </owl:DatatypeProperty>
  <owl:ObjectProperty rdf:ID="taggedBy">
    <rdfs:range rdf:resource="#Vonlunteer" />
    <rdfs:domain rdf:resource="#Tweet" />
  </owl:ObjectProperty>
  <owl:DatatypeProperty rdf:ID="tweetSentBy">
    <rdfs:domain rdf:resource="#Tweet" />
    <rdfs:domain rdf:resource="#Beneficiary" />
  </owl:DatatypeProperty>
  <owl:DatatypeProperty rdf:ID="tweetType">
    <rdfs:range rdf:resource="#TweetCategory" />
    <rdfs:domain rdf:resource="#Tweet" />
  </owl:DatatypeProperty>
  <owl:ObjectProperty rdf:ID="tweetType" />
  <owl:ObjectProperty rdf:ID="volunteerType">
    <rdfs:domain rdf:resource="#Vonlunteer" />
    <rdfs:range rdf:resource="#AuthenticatedVolunteer" />
    <rdfs:range rdf:resource="#AnonymousVolunteer" />
  </owl:ObjectProperty>
</rdf:RDF>
```

Figure 77: Humanitarian Logistic Ontology - OWL Code

APPENDIX B – PROTOTYPE IMPLEMENTATION AND USER GUIDE

The aDApTA implementation is composed of the ETL4LOD-Graph framework and the application prototype. As explored in Section 4.3.2, the ETL4LOD-Graph is a set of steps which are plugins to the PDI (Pentaho Data Integrator) framework. They are ready to be applied in any domain. As explored in see Chapter 5, the prototype, applied in the Humanitarian Logistic case, is composed of the data sources, the ETL workflows, the corresponding supporting files, the domain information base endpoint, and a set of online reports. These components are all necessary to reproduce the implementation used by aDApTA.

The ETL4LOD-Graph steps code files are available at GitHub (https://github.com/Kellli/ETL4LOD-Graph). It was developed with Java language (jdk1.6.0_26[38]) and the Apache Maven (apache-maven-3.2.1[39]) to manage the project building and deployment. The prototype files are available on the research group site, at the aDApTA web page (http://greco.ppgi.ufrj.br/lodbr/index.php/principal/adapta). The screen shots of these web pages are illustrated in Figure 78 and Figure 79.

The first step to use the ETL4LOD-Graph is installing the version 4.3.0-stable[40] of PDI. After that, download the step files and move them to the plugin directory, as illustrated in Figure 80. Thereafter, when the PDI is opened, the ETL4LOD-Graph steps (Figure 81) will be available to be used and applied in any domain. The steps of ETL4LOD (Figure 82) will also be available. Furthermore, the job of the ETL4LinkedProv (see Section 3.5) is available at https://github.com/rogersmendonca/provenance_collector. Once installed, following the same procedure of the steps installation, the job (Figure 83) will be available.

Thereafter, to run the prototype, download the files of data sources, ETL workflows, and corresponding supporting files. They can be installed in any directory which can be read by PDI. However, it is necessary to set the new path on the steps that use the files. For convenience, it was used environment variables as illustrated in Figure 84.

---

[38] http://www.oracle.com/technetwork/java/javase/downloads/java-archive-downloads-javase6-419409.html#jdk-6u26-oth-JPR

[39] http://maven.apache.org/download.cgi

[40] http://sourceforge.net/projects/pentaho/files/Data%20Integration/4.3.0-stable/

The domain information base endpoint can be configured downloading and installing the Virtuoso Server[41]. The prototype endpoint is available at http://crbd.ppgi.ufrj.br:8890/sparql and http://crbd.ppgi.ufrj.br:8890/conductor, which can be accessed through a query and administrator interface, respectively, as illustrated in Figure 85 and Figure 86. The set of online reports are available through live examples at http://greco.ppgi.ufrj.br/lodbr/index.php/adapta/#LiveReports, as illustrated in Figure 87.

The main package is the *R-aDApTA Phillipines – Graph* job (Figure 42) which call all other packages. To run this main job, click the button play, highlighted in the Figure 88. However, to collect the corresponding provenance, open the *job_provenance.kjb* file. It is set to run the *R-aDApTA Phillipines – Graph* job and to collect the corresponding provenance data, as illustrated in Figure 89.

After running the workflow packages, some online reports were built using the SGVIZLER[42] and the YASGUI[43] APIs. The source code of the reports can be accessed in the project webpage. An example of an HTML code is illustrated in Figure 90.

[41] http://virtuoso.openlinksw.com/

[42] http://dev.data2000.no/sgvizler/

[43] http://doc.yasgui.org/

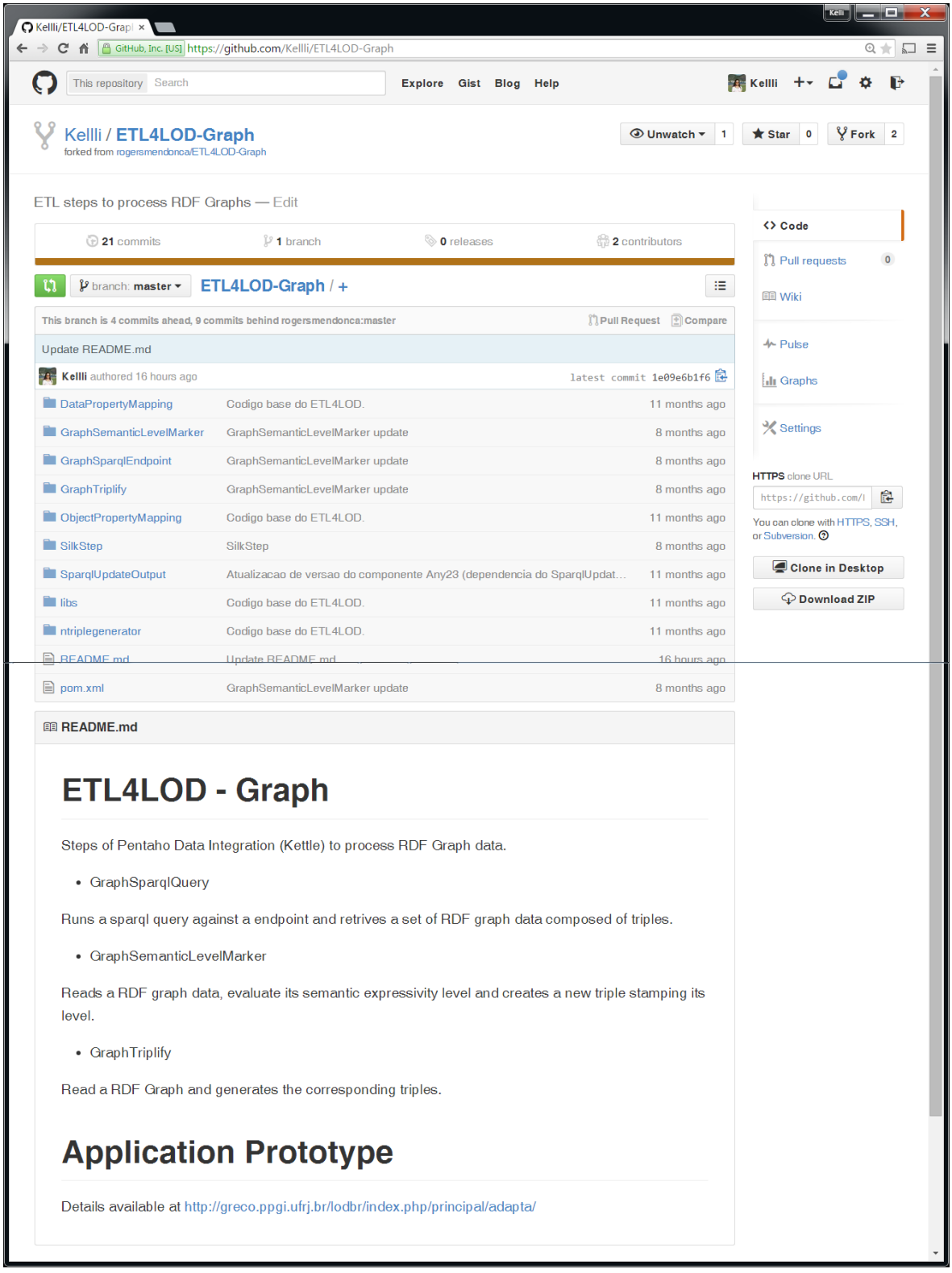


Figure 78: ETL4LOD-Graph web page at GitHub site

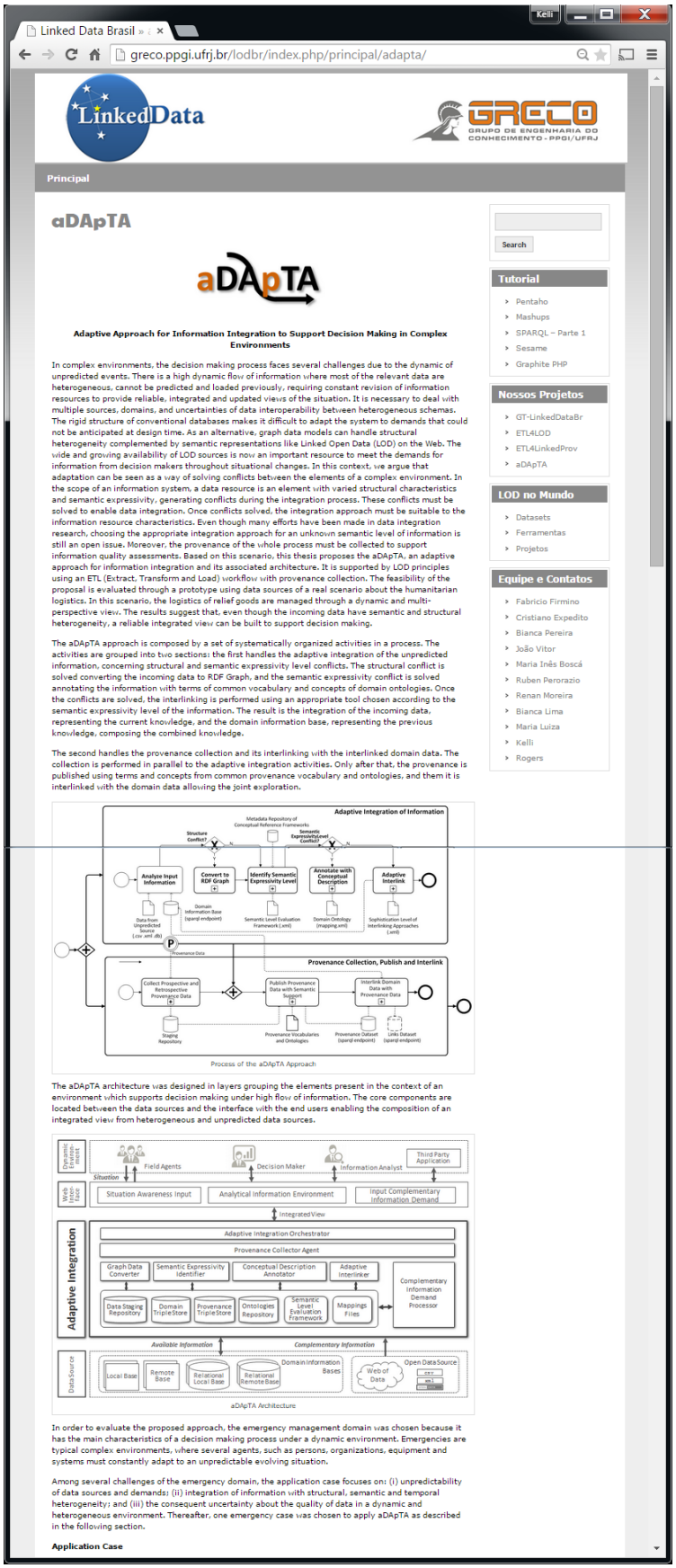


Figure 79: aDApTA web page

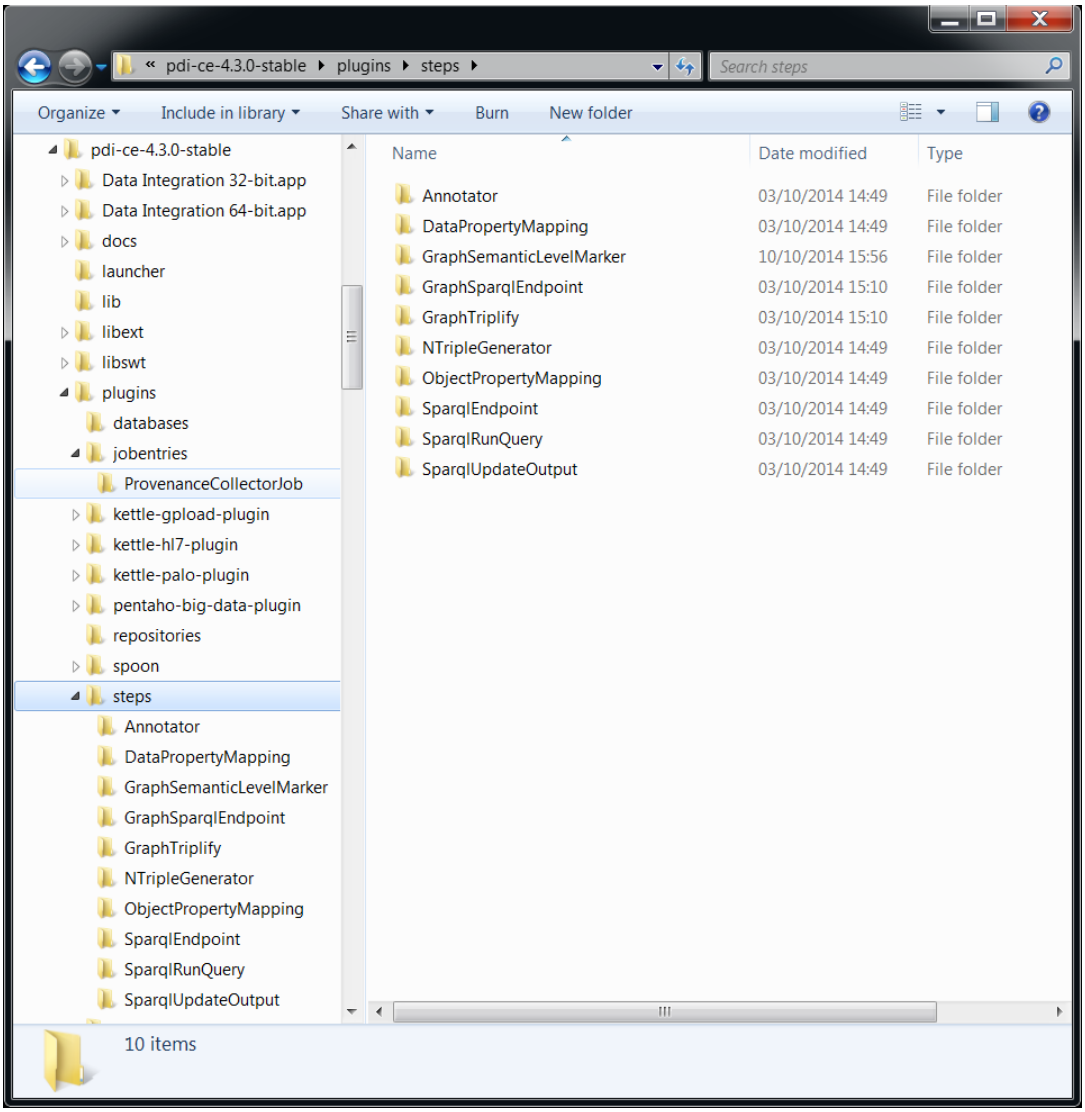


Figure 80: Tree path directory of PDI steps pluging

Icon	Step Name	Description
	RDF Graph Semantic Level Marker	Reads a triple, evaluate its semantic expressivity level and creates new triple stamping its level.
	RDF Graph Sparql Query	Runs a sparql query against a endpoint and retrives a set of RDF graph data composed of triples.
	Triple Annotator	Annotates a triple with terms form vocabularies and ontologies according to a mapping from-to text file.
	RDF Graph Triply	Receives na RDF Graph and deliver three fields (subject, predicate, object) for each triple of the graph.

Figure 81: ETL4LOD-Graph steps catalogue

Icon	Step Name	Description
	Sparql Query	Runs a sparql query against a endpoint and retrives a set of triples.
	Sparql Insert	Receives a field with a triple and insert it on a triplestore through a enpoint.
	DataPropertySetting	Subject and Predicate setup of a literal object
	ObjectPropertySetting	Predicate setup of a URI object
	NTriple Generator	Generates a triple from three fields with a subject, a predicate and a object.
	Sparql Run Query	Receives a field with a sparql query and executes it against a sparql endpoint.

Figure 82: ETL4LOD steps catalogue

Icon	Step Name	Description
	Provenance Collector	Job entry that collects retrospective and prospective provenance data during a Kettle Job execution, including the ETL4LOD-Graph steps.

Figure 83: ETL4LinkedProv job catalogue

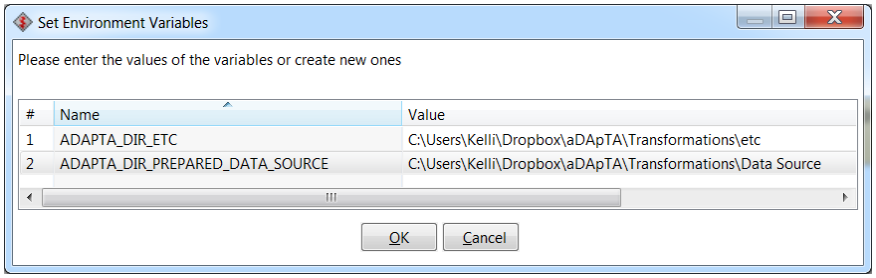


Figure 84: Interface of setting environment variables on PDI

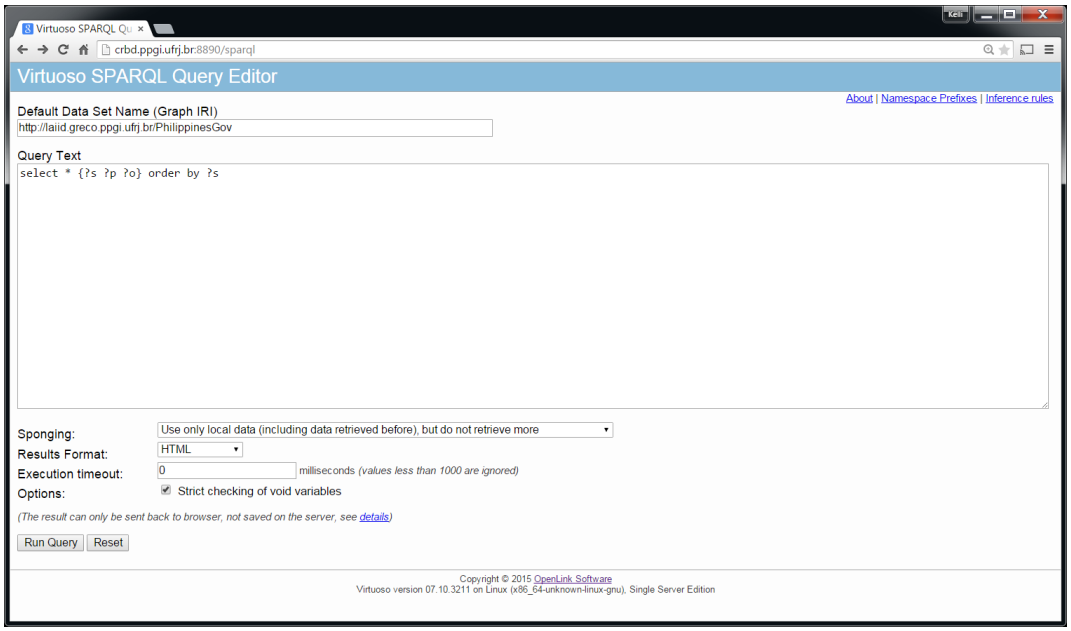


Figure 85: Virtuoso SPARQL query interface of the prototype endpoint

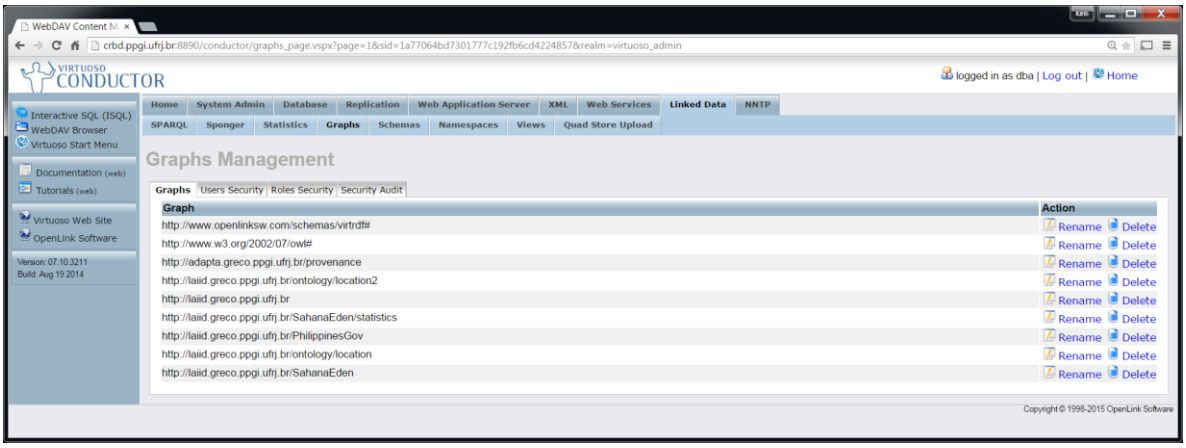


Figure 86: Virtuoso administrator interface of the prototype endpoint

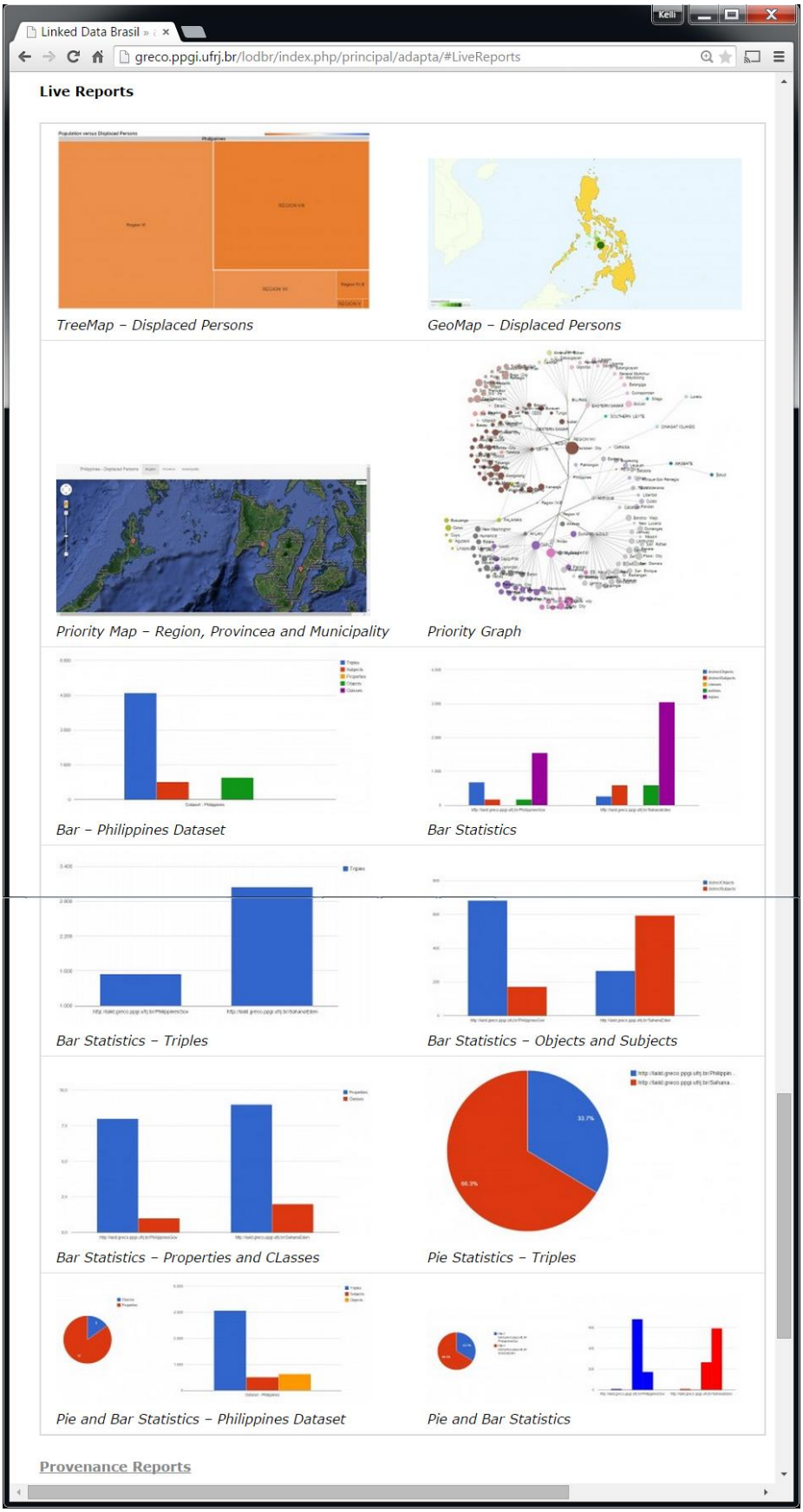


Figure 87: Web page of the prototype live reports

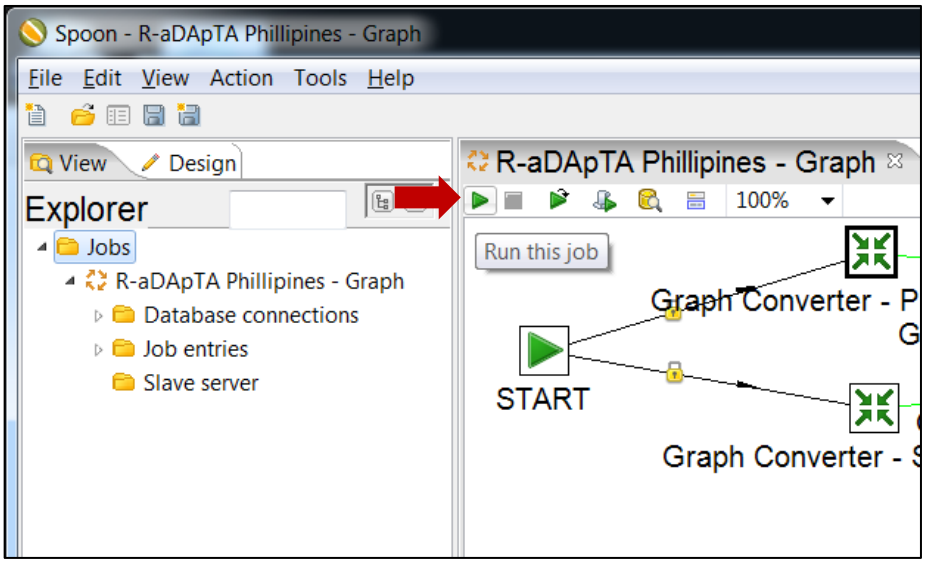


Figure 88: Running a job on PDI

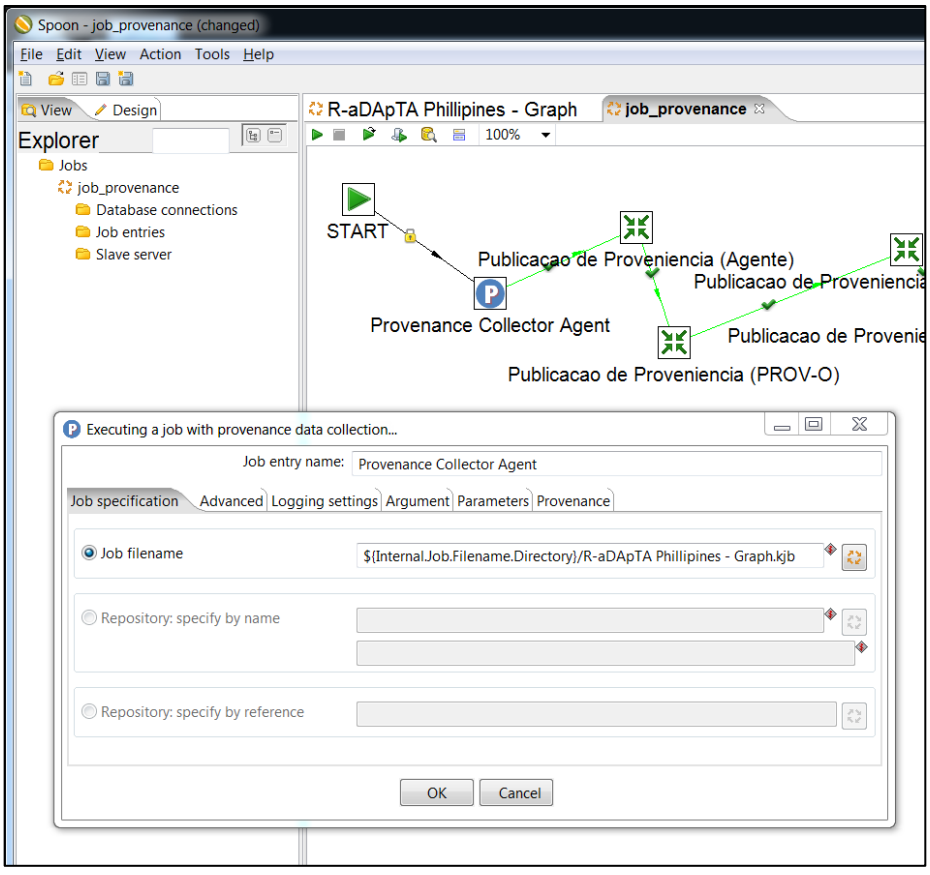


Figure 89: Provenance job parameter

```
<!DOCTYPE HTML>
<html>
     <head>
      <meta http-equiv="Content-Type" content="text/html; charset=utf-8">
      <link rel="stylesheet" type="text/css" href="querybox.css">
      <title>Statistics(Pie Chart)</title>
      <meta charset="UTF-8">
      <script type="text/javascript"
src="http://ajax.googleapis.com/ajax/libs/jquery/1.9.0/jquery.js"></script>
    <script type="text/javascript" src="https://www.google.com/jsapi"></script>
    <script type="text/javascript"
src="http://beta.data2000.no/sgvizler/release/0.6/sgvizler.js"></script>
    <script type="text/javascript">
       sgvizler.prefix('npd', 'http://sws.ifi.uio.no/npd/');
       sgvizler.prefix('npdv', 'http://sws.ifi.uio.no/vocab/npd#');
       $(document).ready(function() { sgvizler.containerDrawAll(); }); </script>
   </head>
  <body>
  <div id="statistics_bars" data-sgvizler-
endpoint="http://crbd.ppgi.ufrj.br:8890/sparql"
    data-sgvizler-query="
    select distinct ?Dataset ?Triples
     from <http://laiid.greco.ppgi.ufrj.br/PhilippinesGov/statistics>
     from named <http://laiid.greco.ppgi.ufrj.br/SahanaEden/statistics>

    {
        {
        ?a  <http://www.w3.org/2000/01/rdf-schema#seeAlso> ?Dataset ;
        <http://rdfs.org/ns/void#classes> ?classes ;
        <http://rdfs.org/ns/void#distinctObjects> ?distinctObjects ;
        <http://rdfs.org/ns/void#distinctSubjects> ?distinctSubjects ;
        <http://rdfs.org/ns/void#entities> ?entities ;
        <http://rdfs.org/ns/void#properties> ?properties ;
        <http://rdfs.org/ns/void#triples> ?Triples.
        }
        UNION
         { graph ?g
            {
        ?a  <http://www.w3.org/2000/01/rdf-schema#seeAlso> ?Dataset ;
        <http://rdfs.org/ns/void#classes> ?classes ;
        <http://rdfs.org/ns/void#distinctObjects> ?distinctObjects ;
        <http://rdfs.org/ns/void#distinctSubjects> ?distinctSubjects ;
        <http://rdfs.org/ns/void#entities> ?entities ;
        <http://rdfs.org/ns/void#properties> ?properties ;
        <http://rdfs.org/ns/void#triples> ?Triples.
            }
        }
    }"
    data-sgvizler-chart="google.visualization.PieChart"
    data-sgvizler-loglevel="2" style="width:1000px; height:500px;"></div> </div>
  <div id="YASGUI">
    <iframe src="OMITTED FOR LEGIBILITY. SEE LIVE REPORT."
    style="border:0; width:1320px; height:550px; margin-top:20px"></iframe>
    <noscript>
        <div style="width: 22em; position: absolute; left: 50%; margin-left: -11em;
color: red; background-color: white; border: 1px solid red; padding: 4px; font-
family: sans-serif"> Your web browser must have JavaScript enabled in order for
this application to display correctly. </div>
    </noscript>
   </p>
  </div>
  </body>
</html>
```

Figure 90: An example of HTML code of a prototype live report