Technology
Arts Sciences
TH Köln

# Improving e-commerce fraud investigations in virtual, inter-institutional teams:

## Towards an approach based on Semantic Web technologies

Master Thesis

by

Andreas Gerlach

submitted to obtain the degree of

Master of Science (M.Sc.)

at

TH Köln - University of Applied Sciences
Institute of Informatics

Course of Studies

Web Science

First supervisor: Prof. Dr. Kristian Fischer
TH Köln - University of Applied Sciences

Second supervisor: Stephan Pavlovic
TH Köln - University of Applied Sciences

Cologne, August 2016

**Contact details:** Andreas Gerlach
Wilhelmstr. 78
52070 Aachen
andreas.gerlach@smail.th-koeln.de

Prof. Dr. Kristian Fischer
TH Köln - University of Applied Sciences
Institute of Informatics
Steinmüllerallee 1
51643 Gummersbach
kristian.fischer@th-koeln.de

Stephan Pavlovic
TH Köln - University of Applied Sciences
Institute of Informatics
Steinmüllerallee 1
51643 Gummersbach
stephan@railslove.com

# Abstract

There is a dramatic shift in credit card fraud from the offline to the online world. Large online retailers have tried to establish countermeasures and transaction data analysis technologies to lower the rate of fraudulent transactions to a manageable amount. But as retailers will always have to make a trade-off between the *performance* of the transaction processing, the *usability* of the web shop, and the overall *security* of it, one can assume that e-commerce fraud will still happen in the future. Thus, retailers have to collaborate with relevant business partners on the incident to find a common ground and take coordinated (legal) actions against it.

Trying to combine the information from different stakeholders will face issues due to different wordings and data formats, competing incentives of the stakeholders to participate on information sharing, as well as possible sharing restrictions that prevent them from making the information available to a larger audience. Moreover, as some of the information might be confidential or business-critical to at least one of the parties involved, a *centralized* system (e.g. a service in the public cloud) can not be used.

This Master Thesis is therefore analysing how far a computer supported collaborative work system based on peer-to-peer communication and Semantic Web technologies can improve the efficiency and effectivity of e-commerce fraud investigations within an inter-institutional team.

**Keywords:** e-commerce fraud investigation, Semantic Web, Shared Information Spaces

# Contents

# 1 Introduction

This introductory chapter of the Master Thesis starts with a section showing the importance and relevance of the topic in the research area of Web Science, which is followed by a short description of the problem this Master Thesis focuses on, and ends with an outline of its structure.

## 1.1 Motivation

> *"When it comes to fraud, 2015 is likely among the riskiest season retailers have ever seen, [...] it is critical that they prepare for a significant uptick in fraud, particularly within e-commerce channels." (Reuters 2015)*

This statement from Mike Braatz, senior vice president of Payment Risk Management, ACI Worldwide in (Reuters 2015) shows the dramatic shift in credit card fraud from the offline to the online world that retailers are starting to face nowadays.

In general, credit card fraud can occur in the physical world if a consumer has lost the credit card, or if the credit card has been stolen by a criminal. This usually results in an identity theft by the criminal, who is using the original credit card to make financial transactions by pretending to be the owner of the credit card. Additionally, consumers might hand over their credit card information to untrustworthy individuals, who might use this information for their own benefit. In the real world scenario there usually is a face-to-face interaction between both parties. A consumer, wanting to do business with a merchant or interacting with an employee of a larger business, has to hand over the credit card information explicitly and can deny doing so in a suspicious situation. The criminals on the other hand must get access to the physical credit card first, before they are able to make an illegal copy of it — a process called skimming. The devices used to read out and duplicate the credit card information are therefore called skimmers. These can be special terminals that the criminals use to make copies of those credit cards they get their hands on, or those devices can be installed in or attached to terminals the consumers interact with on their own (Consumer Action 2009). All of these so-called *card-present transaction* scenarios have seen a lot of improvements in

security over the last years. Especially the transition from magnetic swipe readers to EMV chip-based credit cards makes it more difficult for criminals to counterfeit them (Lewis 2015).

As a consequence criminals are turning away from these card-present transaction scenarios in the offline world. Instead they turn their attention to transactions in the online and mobile world, in which it is easy to pretend to own a certain credit card. Most online transactions (either e-commerce or m-commerce) rely *only* on credit card information such as card number, card holder and security code for the card validation process; therefore these interactions are normally called *card-not-present transactions*. The credit card information can be obtained by a criminal in a number of ways. First they might send out phishing emails to consumers. These emails mimic the look-and-feel of emails from a merchant or bank that the consumers are usually interacting with, but instead navigate them to a malicious web site with the objective to capture credit card or other personal related information (Consumer Action 2009). Additionally, criminals can break into the web sites of large Internet businesses with the intent of getting access to the underlying database of customer information that in some cases also holds credit card data (Holmes 2015). Also, as some of the online retailers do not encrypt the transaction information before transmitting them over the Internet, a hacker can easily start a man-in-the-middle attack to trace these data packages and get access to credit card and personal related information this way (Captain 2015).

Based on these facts it should not come as a surprise that the growth rate of online fraud has been 163% in 2015 alone (PYMNTS 2016). This results in huge losses for the global economy every year and it is expected that retailers are losing $3.08 for every dollar in fraud incurred in 2014 (incl. the costs for handling fraudulent transactions) (Rampton 2015). These fraudulent transactions have also an impact on the revenue of the online retailers, who have seen a growth of 94% in revenue lost in 2015. Overall it is estimated that credit card frauds resulted in losses of $16 billion globally in 2014 (PYMNTS 2016) (Business Wire 2015).

While it is possible to prevent fraudulent transactions in the card-present, real-world scenario (mainly due to the introduction of better technology and the establishment of organizational countermeasures in the past), it is more difficult to do so in the card-not-present e-commerce and m-commerce scenarios, which are lacking face-to-face interactions and enable massive scalability of misusing credit card information in even shorter time frames (Lewis 2015). Large online retailers have tried to establish countermeasures and transaction data analysis technologies to lower the rate of fraudu-

lent transactions to a manageable amount. But this is still an expensive and inefficient solution to integrate into the retailers' business operations, which is largely driven by machine-learning techniques and manual review processes (Brachmann 2015). Moreover, it can be assumed that the online retailers are getting into a "Red Queen's race" with the criminals here. With every new technology or procedure introduced they might just be able to safe the status quo. This is largely due to the facts, that there will be no 100% security for a complex and interconnected system such as an e-commerce or m-commerce shop, the criminals will also increase their efforts and technology skills to adapt to new security features; and most importantly retailers will always have to make a trade-off between the *performance* of the transaction processing, the *usability* of the web shop, and the overall *security* of it.

## 1.2 Problem Definition

This Master Thesis develops a concept to optimize the collaboration between the affected stakeholders in case of an existing credit card fraud in an e-commerce system. It does *not* look into novel techniques and methods to *prevent* credit card fraud in the e-commerce scenario. This aspect has been seeing a lot of research in the last years.[1]

Stakeholders might include vendors and other businesses that a retailer has a long-term business relationship with, law enforcement agencies, payment service providers such as PayPal or Visa, banks, and even competitors, which are also affected by an online fraud. In these cases merchants usually try to solve the issues on their own, and try to get in contact with relevant parties by phone or e-mail if necessary. But these communication styles do not fit to the complexity of the task involved, and based on the media-richness model (see Figure 1.1) will result in inefficient and ineffective problem solutions.

Due to the task complexity a physical face-to-face meeting with representatives of all stakeholders involved might be a good fit, but arranging such a meeting (at the same time and at the same place) with multiple parties, which are globally dispersed, is either economically not feasible or takes a lot of time. But the more time passes for investigating a fraud, the more difficult it will become to identify the fraudsters and take legal actions against them. Acting in a timely fashion can therefore reduce the

[1] Please also note the various US patent applications of Google on that matter from 2015, e.g.: "Credit card fraud prevention system and method", "Financial card fraud alert", "Payment card fraud prevention system and method" (Google Patents).

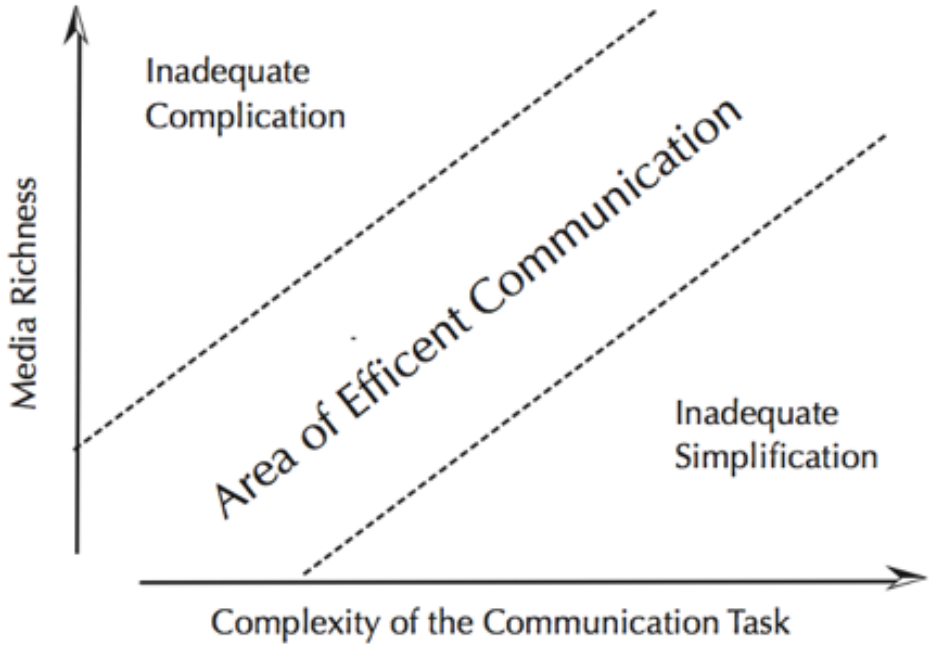


Figure 1.1: The media-richness model (Rice 1992)

risk of losing the money completely.

As of these conditions a computer-supported collaborative work (CSCW) system might be an alternative to *collaborate* on an incident of e-commerce fraud (at the same time, but at different places). CSCW systems can be categorized by their support for the mode of group interaction as done in the "3C model" (Koch 2008):

- **Communication:** two-way exchange of information between different parties,
- **Coordination:** management of shared resources such as meeting rooms,
- **Collaboration:** members of a group work together in a shared environment to reach a goal.

Based on the level of support for one of these functionalities the various systems can be classified and described as shown in Figure 1.2:

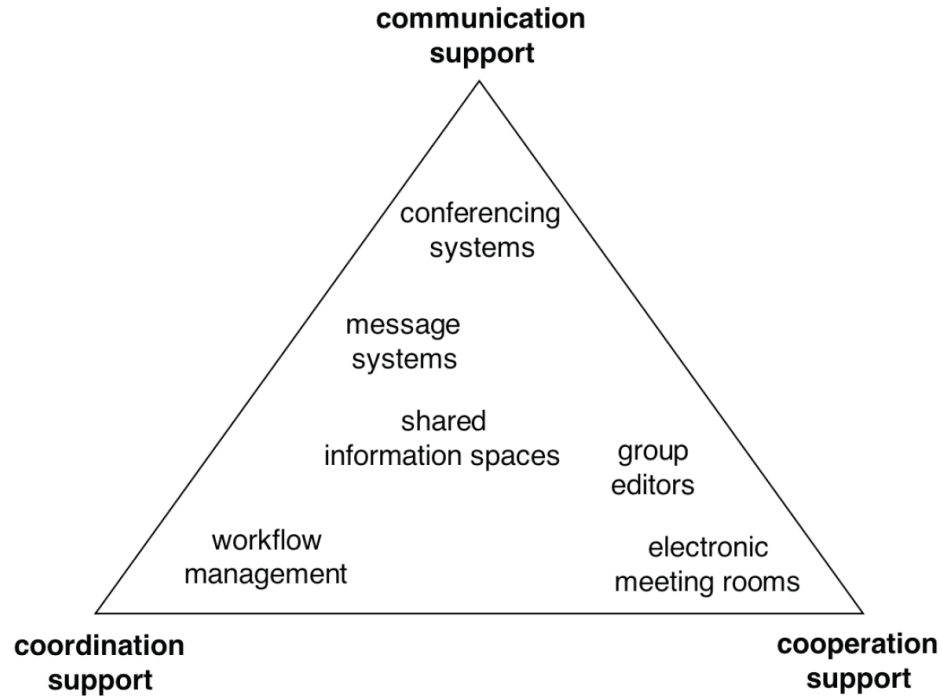


Figure 1.2: The 3C model (Koch 2008)

A good candidate for such a collaborative system *could* be a shared information space; aka team rooms, cloud storage services or document management systems, which allow participating parties to access information at any place, any time, and to share information between each other — usually with a build in versioning support for artefacts and a workflow component.

However, as some of the required information might be confidential or business-critical to one of the involved parties, a centralized system (e.g. a service in the public cloud) can *not* be used in the scenario described here. Another key characteristic of the investigation of an e-commerce fraud is the fact that it involves information sharing from many different organizations. These different aspects have to be combined into a shared information space in a meaningful way to be able to achieve a common group goal on time. Trying to combine information from different stakeholders will face issues due to different wordings and data formats, competing incentives of the stakeholders to participate on information sharing, as well as possible sharing restrictions that prevent making the information available to a larger audience.

Decentralized information sharing architectures, which utilize peer-to-peer communication technologies, are either restricted to a commonly agreed set of data entities and relations between all parties involved, or are lacking richer semantics for sharing and integrating content between the stakeholders. Semantic Web technologies can help lower the barrier to integrate information from various sources into a shared information space, and the advantages of peer-to-peer communication and Semantic Web technologies for information sharing in distributed, inter-organizational settings have been shown in (Staab & Stuckenschmidt 2006).

Still these studies concentrate on making information from different parties searchable and accessible in a distributed, shared information space, in which data can be accessed and queried at any time from any participating party. They are not solving the problem of working collaboratively on a common goal in an ad-hoc, loosely-coupled virtual team of disperse organizations by making certain (sometimes sensitive) information available in a shared environment.

Therefore, the research question for this Master Thesis can be summarized as follows:

> *In how far can a computer supported collaborative work system based on peer-to-peer communication and Semantic Web technologies improve the efficiency and effectivity of e-commerce fraud investigations within an inter-institutional team?*

## 1.3 Outline of this Master Thesis

Before looking into the investigation of e-commerce fraud incidents and their possible examinations, the Master Thesis starts with a description of related works in Chapter 2. These research papers have been evaluated during the course of this Master Thesis, and have had an influence on the concept and design of the collaborative system being developed.

In the next part, Context Analysis in Chapter 3, the Master Thesis discusses the e-commerce scenario in detail. It starts with a description of the e-commerce shopping process, looks into the stakeholders involved, as well as shows possible kinds of e-commerce fraud incidents and how they are handled today. Based on these findings this chapter closes with a presentation of the specific scenario that has been selected for further examination within this Master Thesis.

After this initial scope setup the Master Thesis briefly outlines the theoretical foundations in Chapter 4, which are required for the understanding of the concepts and design decisions in Chapter 5 and Chapter 6. This section starts with a short overview of the relevant facets of computer-supported collaborative work systems (CSCW), shows the essential specifications of the Semantic Web, and ends up with an introduction to the peer-to-peer (P2P) communication techniques and protocols.

The main parts of this Master Thesis (Chapter 5 and Chapter 6) discuss the concept and design for a collaborative system that supports the investigation of e-commerce fraud incidents. These chapters will elaborate and analyse the possibilities for designing and using such a collaborative system. The objective is to come up with an approach at the end of the discussions that might be the best fit for the problem described in the scenario in Chapter 3.

To conclude the Master Thesis also sums up the findings and gives an outlook for future work on this topic.

# 2 Related Works

The study of Pritikana Sen et al. starts with an introduction to the subject of e-commerce and specifies possible types of it. It further shows the benefits of e-commerce (e.g. the global reach of Web shops) as well as its limitations. Here it mentions explicitly the security of the system and the communication protocols used. The paper lists the relevant stakeholders of an e-commerce transaction and describes the credit card payment process in detail. It concludes with an analysis of the security features of a Web shop and shows that those are not limited to technical aspects alone, but always include the consumers and their behaviours on the Internet (Sen et al. 2015).

Sobko's paper defines non-cash transactions and shows ways in which fraudsters try to cheat the system. It starts with a classification of non-cash payments including credit and debit cards that are handed out to individuals by financial institutions. The author also notes possible ways to trick an individual with the objective to get access to credit card information such as phishing and skimming. He explains that once a transaction has been successfully executed with a stolen credit card, the information about it will be sold on the black market to other fraudsters, who will then use the same credit card to make additional purchases. Further on, the paper discusses the impact of fraudulent transactions for the merchants and credit card owners, as well as shows technological advances and regulations that have been developed to protect them against non-cash frauds (Sobko 2014).

The research of Priya J. Rana et al. shows possible frauds in e-commerce and how they can be detected with current fraud prevention systems. They explain different implementations of fraud detection algorithms, which range from simple rule-based filtering to score-based solutions using fuzzy logic. They conclude that actual systems in use can cover up to 80% of fraudulent transactions at manageable efforts and costs. More coverage can be achieved by combining existing solutions with information of the card owners' profile, which would introduce credit card usages patterns into the analysis. Still this latter solution is very expensive to implement and operate (Rana & Baria 2015).

The paper by Carvalho et al. looks into the financial crime investigation process by using banking frauds as an example. It shows that the investigation of them is a very complex task that needs further collaboration between experts. Still just sharing the information will not be sufficient as a common understanding of the different aspects and terms is required. Therefore they state that finding a common language to exchange information is very important for the success of the investigation. Based on this finding the paper also develops an ontology to describe the domain of banking fraud investigation. It elaborates on the objects and their relations in detail and shows that reusing concepts and terms from existing vocabularies can be helpful when designing an ontology of one's own. The authors conclude that semantic technologies can have a positive impact on the crime investigation as they are providing basic reasoning capabilities on the data sets, as well as offering support for the combination of information coming from different sources. Finally they consider semantic technology to be very important in future cybercrime inspections (Carvalho et al.).

The seminal paper "Linked data-the story so far" explains the fundamental concepts, approaches and technologies to share data on the Web. It shows how a RDF data set should be used to publish structured data on the Web, and provide rules how these resources should be described. Additionally, it discusses the commonly used vocabularies available on the Web, and explains ways how to link together various resources on the Internet. After describing different kinds of applications that are possible with the technologies shown, it gives an outlook of future research, which also states the aspects of schema mapping and data fusion as possible challenges. Other issues in the area of Linked data are related to privacy violations that become possible when information from different sources is combined. As conclusion the authors consider the Linked data approach as intermediate step to a Semantic Web, because it follows the same established Web standards such as RDF, RDFS and SPARQL, but uses a more pragmatic approach by getting rid of all the complexities involved when having to create, maintain and use elaborate ontologies in OWL (Bizer et al. 2009).

Different ways to publish semantic data on the Web have been analysed by Laurens Rietveld et al. They show that current approaches range from simple data dumps of complete RDF data sets to publicly available query endpoints using the SPARQL protocol and query language. As a conclusion they state that the more flexible approach of SPARQL endpoints is generally not working on the Internet, but is instead more suitable for internal data collection and analysis. This is due to the enormous overhead of a SPARQL server both in memory and CPU consumption if it has to deal with a large RDF data set and a huge number of concurrent users. The proposal, which has

been developed in their research paper, uses novel approaches called “Linked Data documents” and “Triple Pattern Fragments”. The former provides subsets of a RDF data set optimized for specific subjects or objects and allow focused querying of a RDF data set. The latter tries to move parts of the processing of a large scale RDF data set from the server to the clients and builds upon the ideas of the “Linked Data documents”. As a result of the paper new approaches are proposed for offering RDF data sets on the Web, which are optimized for querying and processing large scale RDF data sets, and make better use of the processing capabilities of clients (Rietveld et al. 2015).

The need for a RDF vocabulary to express products and offerings on the Internet was first mentioned and described by Martin Hepp. The author is also the founder of the GoodRelations vocabulary, which he explains in detail in his paper. After showing possible use case scenarios for such an ontology on the Web the author elaborates the available entities and their meanings. Interestingly, the author states that he has restricted usage of more expressive OWL axioms in the GoodRelations vocabulary due to the limited availability of full-featured OWL reasoners. By focusing on RDFS constraints only less advanced functionality is required for the processing engine of the RDF data sets. The paper closes with examples of using the vocabulary in the e-commerce scenario and its possible future development for B2B service integrations (Hepp 2008).

With the wide adoption of semantic data on the Web there was a growing need to harmonise vocabularies used to express commonly shared objects and entities such as events, products, places, and people on the Internet. Even though there were initial approaches from the Semantic Web community such as FOAF for people, those did not cover every aspect leading search engine providers were looking at to enhance their search results. Due to these circumstances they started an initiative of their own to express and define a meta data vocabulary that covers a wide range of topics discussed on the Web. This initiative resulted in the Schema.org specifications, which due to the influence on search engines such as Google Search, Microsoft Bing or Yahoo! has seen wide adoption by businesses and individuals on the Web. The history of and vision behind this initiative is depicted in the paper from R.V. Guha et al. They also show how the initial idea to embed meta data directly into the HTML of a Web site leads to further development of the RDFa syntax. The paper closes with an outlook on the future development of Schema.org including the available extension mechanism, which has already been used by certain verticals such as automotive and healthcare (Guha et al. 2016).

The research paper of Christian Vogt et al. analyses in how far WebRTC data channels can be used to build a large structured P2P communication network and exchange arbitrary data on it. The authors describe a multi-layered approach to establish P2P connections and route messages through the network. All of their work is based on the WebRTC standard from the W3C without requiring additional software on the client side (beside a Web browser with WebRTC support). They choose a structured P2P network architecture based on a distributed hash-table to speed up managing of and searching for information in the network. They conclude that WebRTC as open standard for distributed P2P messaging shows a lot of potential for new kinds of applications, and that the standard will need deeper exploration of its potentialities. Additionally, they mention the aspects of security and privacy that have to be looked at in detail, but also noticed that there are already groups in the standard councils working on topics such as end-to-end encryption of messages and authentication of peers (Vogt et al. 2013b).

In another research paper Christian Vogt et al. show how to build a distributed content-publishing platform on WebRTC named BOPlish. Their approach does not rely on any Web server to host and provide access to user generated content, but works by connecting Web browsers via WebRTC data channels directly. In this virtual P2P content network anyone can easily publish and retrieve information. The proposed application also includes security and authentication features based on the latest WebRTC standards, which allow to securely share information between the peers in this P2P network. Their proposal still relies on a bootstrap server to setup and connect new peers to the network, as well as a distributed hash-table to index and look up available information. A novel approach in the paper is to use a custom naming and addressing schema to access and retrieve the information in the P2P network. Peers will be authenticated and assigned to a section in a distributed hash-table by the bootstrap server. Information available on a peer can further be addressed by the custom addressing schema directly, which follows the well-known URI schema but uses the identity of the owner of the information to address them. When accessing information with this addressing schema the identity of the owner has to be resolved first, before the information can be retrieved from the actual peer holding them. They conclude that this approach offers decentralized content-publishing between browsers and enables users to share information without the need for a Web server. By providing a new addressing schema for the content available on the network, which makes accessing the information independent of its current IP address, their approach is better suited for the dynamics of large P2P networks (Vogt et al. 2013a).

Robert Pienta et al. show in their research paper that understanding and making sense out of graph-formatted data is a complex and challenging process, which involves bringing together findings from different research fields such as data mining, machine learning, human computer interaction, and information visualisation. They explain that many of today's data sets can be represented as graphs, and making sense out of them requires specialized tools that allow users to interactively explore, visualise, and understand large scale graphs. The authors differentiate between two methods to build an understanding of the information within a graph: top-down or global views and bottom-up or local views. The former procedure starts with an overview of the graph as a whole and allows users to filter and zoom in to areas of interest. The latter one focusses on certain aspects of the information and allows users to navigate and analyse data that relates to these aspects. Thus, tools for analysing graph data have to provide different views to support the sensemaking of the information in the graph. Further on, the authors classify the existing techniques based on these two paradigms and explain each of them briefly. Clustering of a graph, which refers to a method for organising nodes with similar attributes and grouping them together on a global view, can be accomplished with the existing k-SNAP algorithm that also supports drill-down and roll-up operations on the information. For the exploration of a graph, which describes the interactive investigation of it with the intent to develop new insights, and which is part of the bottom-up approaches of graph investigations, the system can use data mining and machine learning techniques to support the user in the sensemaking process. Especially the available methods for discovering common subgraphs in a larger graph can be helpful in situations that try to find abnormal activities in a network such as frauds (Pienta et al. 2015).

The paper of Olivia Angiuli et al. looks into the subject of privacy issues that come up when large amounts of data about humans are gathered, analysed, and eventually shared with others. Due to the fact that analysing and sharing information about humans is subject to laws and regulations, which are specifically designed to protect them, the data has to be anonymised or de-identified prior to making them publicly available. This process should remove certain information from the data set so that individuals can no longer be identified. It usually starts with an identification of the information that needs to be protected. To de-identify the information in a data set the values can be either generalised into broader categories or contexts, or suppressed from the shared data set completely. They conclude that the process of anonymising data is a complex task that has to take care of any biases, which might be introduced during the process, and also has to prevent the shared information from subsequent re-identification of individuals (Angiuli et al. 2015).

# 3 Context Analysis

This chapter looks into the scenario of e-commerce fraud investigation in detail. It starts with an in-depth description of the e-commerce scenario followed by an analysis of the stakeholders involved. It further describes the kind of information each stakeholder has in their local context, and their objectives to take part on the information sharing and collaboration initiative. Based on the analysis of the possible kinds of e-commerce fraud incidents and the current process of their investigation, the chapter closes with a description of the specific scenario, which has been selected for this Master Thesis.

## 3.1 An overview of e-commerce

*E-commerce* as a term relates to the trading of products or services utilizing a computer network such as the Internet. It is usually divided into the following four different subfields (Sen et al. 2015):

1. **Business-To-Business (B2B)**: refers to electronic trading between companies with the objective to improve their supply chain processes,
2. **Business-To-Consumer (B2C)**: refers to electronic trading between a company and its consumers (most prominent example for it is Amazon (Amazon.com)),
3. **Consumer-To-Consumer (C2C)**: refers to electronic trading between consumers (most publicly known example for that is eBay (eBay Inc)),
4. **Consumer-To-Business (C2B)**: refers to electronic trading between consumers and businesses (most notable example for this is TaskRabbit (TaskRabbit)).

Due to the problem initially sketched out in Section 1.1 this Master Thesis will *solely* focus on the B2C aspect of e-commerce. In that case a consumer uses an e-commerce shop of a merchant on the Internet to order products or services online. The merchant offers a catalogue of available products or services on the Web that is available and accessible by the general public and usually has a nation-wide if not global reach.

The merchant can either run the e-commerce shop software on their own servers (on-premise) or can outsource this additional sales channel to a $3^{rd}$ party hosting company or Cloud Service Provider (CSP). Also, the e-commerce shop software itself can be either developed by the merchant in-house or acquired as a boxed product from an Independent Software Vendor (ISV) on the market. For business accounting purposes the merchant also runs a bank account with an acquirer (see Figure 3.1).

When placing an order with a merchant online, the consumers normally use a credit card for finalizing the transaction. These credit cards have originally been handed out to the consumers by the issuers. Additionally, in some online shops it is mandatory for the consumers to create a user account with them, while in others it is not. The former is the preferred way when consumers are repetitively buying from that merchant, whereas the latter might be used for one-time or irregular shopping trips online. To be able to connect to the Internet the consumers also rely on a service offered by an Internet Service Provider (ISP). The whole initial setup for participating in e-commerce activities is found in Figure 3.1.

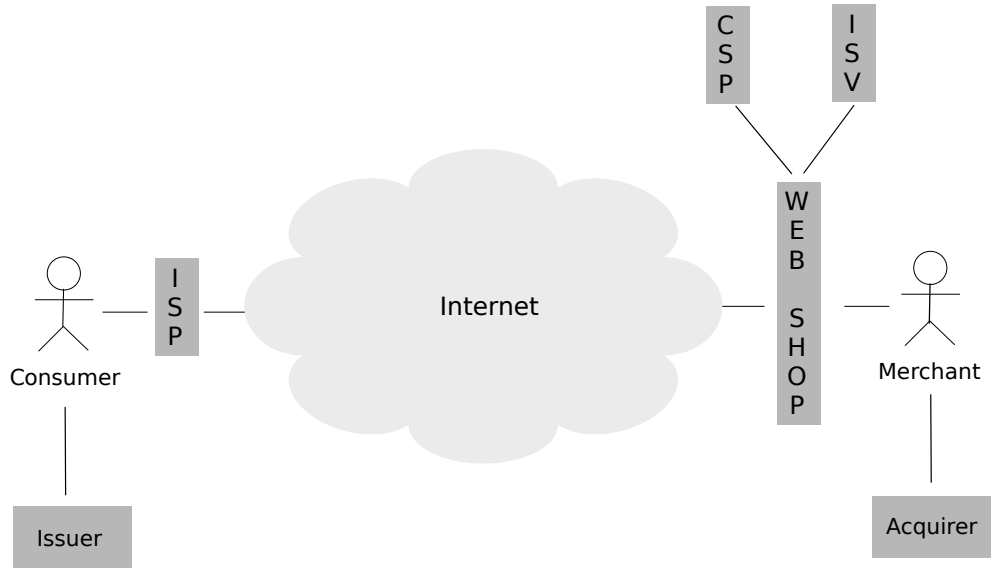


Figure 3.1: E-commerce fundamentals

When a consumer places an order online, the merchant receives at least a list of products or services from the current shopping cart of the consumer, the identification of the consumer, as well as the delivery address to ship the physical items to. If the transaction is going to be finalized with a credit card, the consumer will have to provide additional information such as the billing address and the credit card details (including

the number, the expiry date and the security code of the card).

The merchants usually do not validate the credit card information on their own. For that purpose they are relying on another $3^{rd}$ party service offered on the Internet by the Payment Service Provider (PSP). These providers either validate the credit card information themselves based on a user profile the consumer has with the PSP (e.g. a globally available Web service such as PayPal), or communicate with the issuer of the credit card for doing so. For initiating this validation process the merchant hands over the billing information to the PSP, which also include the credit card details given by the consumer.

Either the PSPs or the issuers validate the correctness of these information with reference to criteria such as:

- Does the billing address match the current consumer's postal address on file?
- Is the stated credit card information correct?
- Is the credit card still valid?
- Is the credit card not marked as being blocked in the internal databases?

The merchant receives the status of the authorization as well as a unique payment token in return. If the authorization has been successful, the merchant collects the items and sends out a shipping request to one of the available Logistic Service Providers (LSP), which are capable of delivering the order. They pickup the items at the merchant's facility and ship them to the delivery address stated by the consumer. Usually at about the same time the merchant informs the acquirer about the order, amount due as well as the payment token received from the PSP. The acquirer is in charge to withdraw the amount of the order from the consumer's bank account either via the PSP or directly from the issuer, depending on who of them has authorized the initial payment request — a process called clearing (Visa Europe 2014). The sequence of activities within an e-commerce checkout process is visualized in Figure 3.2.

## 3.2 Stakeholders

The following section looks at each stakeholder involved in the e-commerce scenario in detail, lists the kind of information they own or provide to others, as well as describes the role of each stakeholder in the e-commerce fraud investigation process (if any).

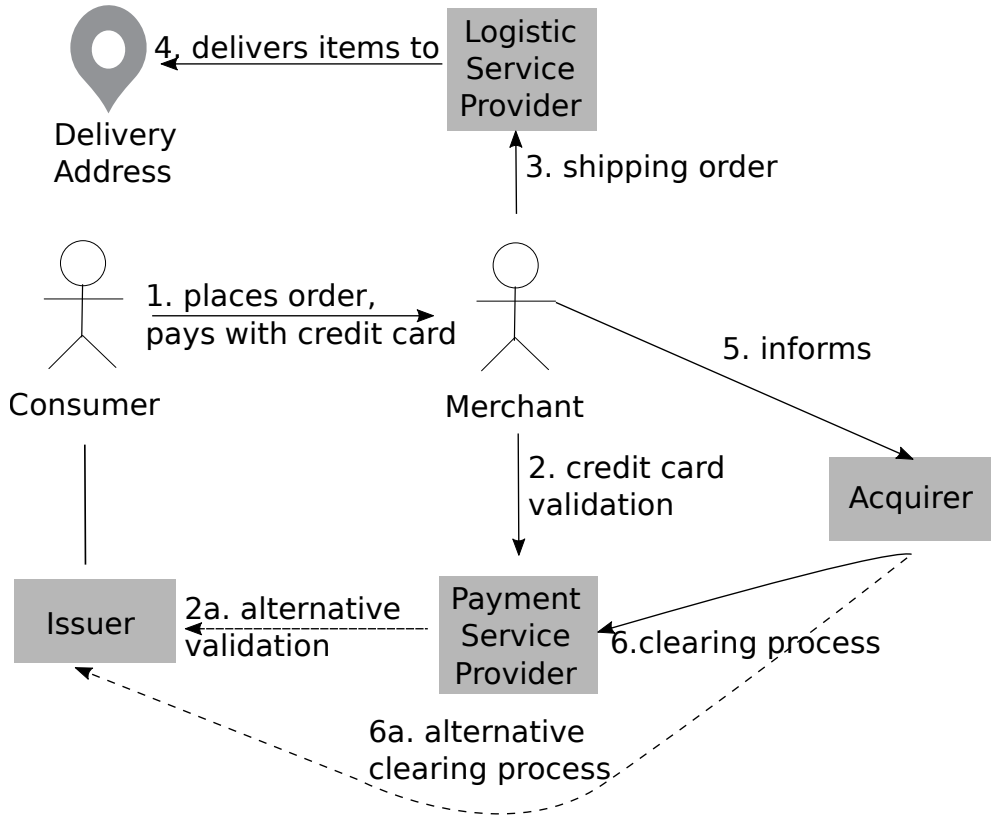


Figure 3.2: E-commerce checkout process in detail

### 3.2.1 Consumer

The consumers are the initiators of e-commerce transactions. They use the shop of a merchant on the Internet to order products or services. For doing so they have to know the URL of the Web shop, have to be connected to the Internet via an ISP, and have to use a standard software called a Web browser on their computer. For the duration of their online browsing sessions they also own a unique IP address provided by the ISP.

They might have had a long-term business relationship with the merchant and already own an user account on the Web shop. As an alternative they might be just interested into a one-time shopping trip, and might want to order the items without creating an account first — sometimes also called “anonymous” or “guest” checkout in the e-commerce shops.

The consumers also have a bank account, and at least own a debit card from that bank to get access to the money on the account. In addition to that they can also hold multiple credit cards. A credit card can be issued by the same bank, or can be

provided by another financial service institution (e.g. American Express). In any case the organization that has handed out the credit card to the consumer is called the issuer.

If the consumers are going to order items in a Web shop, they will usually browse the product and service offerings of a merchant and put the items of interest into the shopping cart first. During the completion of the transaction they have to state the following information to the merchant:

- personal information incl. given name, family name and date of birth,
- the address the items should be shipped to,
- payment information incl. type of payment and billing address (if different to shipping address).

If they are going to end the transaction with a payment of type credit card, they will also have to provide specific information of the credit card that should be used as payment:

- the owner of the credit card (if it is not belonging to themselves),
- the unique credit card number,
- the expiry date of the credit card (in format MM/YY),
- the security code of the credit card.

The consumers have a special role in the whole scenario. As the online merchants have to deal with the consumers without any face-to-face or real-world interactions, the consumers are also the least trustworthy participants from the point of view of the merchants. As Section 3.4 will show, the consumers are the main subjects of investigation in the case of an e-commerce fraud incident. Therefore, the consumers are not taking any active part in the fraud investigation process.

### 3.2.2 Merchant

The merchants offer products and services on the Internet to the general public. They might use the Internet as an additional sales channel, or rely on it solely for making any business. To provide access to the Web shop a merchant has to register a domain name and a URL with a local domain name registry. This specific URL refers to a fixed public IP address, which the server that runs the Web shop software uses. Normally

the merchants do not operate the servers themselves, but rely on a service offered by a Cloud Service Provider for that purpose. Moreover, the Web shop software itself is usually not provided by the merchants, but bought from an ISV on the market. In any case the merchants have special responsibilities in the Web shop, because they have to take care to configure the products, prices, promotions, payment, and shipment services available. In addition products can be categorized by them into categories and sub-categories for easier navigating and searching the offerings in the Web shop by the consumers later.

The merchants can decide whether they restrict ordering of products to registered users only, or allow anonymous users too. The main benefit of the former is the possibility to analyse the shopping behaviour of individual consumers, whereas the latter will open the business for a wider range of consumers as it includes also those, who do not want to register with any existing online shop. Nevertheless, any consumer activity on the online shop is tracked in the analytic databases of a merchant. This includes not only the items that have been placed into the shopping cart, but also any product that a consumer has looked at during a shopping session. Even though these detailed analytic capabilities are actually synonymous for their usage in target-related advertising, they can also help to decide whether a consumer behaves normally or not within a Web shop.

Any business transaction that a consumer makes with a merchant is stored in the merchant's database. The transaction information contains, but is not limited to:

- personal related information of the consumer,
- the address the items will be shipped to,
- a collection of products with quantities and prices,
- the total amount of the order considering promotions, taxes and fees,
- the selected payment information.

If a consumer wants to pay with credit card, the payment process is not handled by the merchants themselves, but is routed to a Payment Service Provider (PSP) on the Internet instead. To initiate the credit card authorization a merchant is sending a request with the following information to the Web service endpoint of a PSP:

- consumer's billing address,
- given credit card number, expiry date and security code,

- identification of the merchant,
- final amount of the current transaction.

In return of the payment authorization a merchant receives and stores these payment-related information for the transaction:

- the type of credit card used (e.g. Visa, MasterCard, American Express, ...),
- the name of the credit card owner,
- the unique payment token received by the PSP,
- the timestamps and result code of the authorization,
- the authority, who has approved the payment (if the merchant works with multiple Payment Service Providers).

As the merchants will collect a lot of personal and payment-related information over time, they are also one of the major sources of possible data leaks in the e-commerce scenario. Due to this circumstance the Payment Card Initiative, a group of banks, issuers and PSPs, provides rules and guidelines (aka PCI/DSS standards) for securely handling these kind of information in an IT system (Virtue 2009).

The merchants are one of the main actors in the fraud investigation process. They are highly interested in figuring out whether a consumer's transaction is valid or not. This is due to the fact that in case of an e-commerce fraud incident the merchants will likely have to cover the costs (see Section 3.4). Also the online merchant's reputation will suffer, if private information from their databases gets leaked. If a merchant falls victim to fraud incidents multiple times, the economic damages can finally result in a bankruptcy of that merchant.

### 3.2.3 Payment Service Provider

The Payment Service Providers offer payment-related services to online merchants. To be able to do this a PSP provides a Web service interface, which the merchants have to communicate with by sending payment authorization requests to it (see above). The PSPs might be able to authorize the payment requests on their own, or might have to route each of them to the corresponding issuer of the credit card used. For the former procedure the PSPs have to run a database of registered users with credit card information on their own (e.g. a Web service such as PayPal). For the latter they will have to know the issuer of the credit card used, and have to call into the Web service

of that issuer for validation purposes. For verifying a credit card and authorizing the payment a merchant hands over the following:

- credit card owner incl. billing address given,
- credit card number,
- credit card expiry date,
- credit card security code,
- identification of the merchant,
- total amount of the current transaction.

In case the PSPs authorize the payment requests themselves, they will have to securely process the information and return the validation results to the merchants. Each result message also contains a unique payment token that a merchant can refer to later to initiate the clearing process. As of this, the PSPs have to persist the credit card and payment-related information in their databases. According to industry standards they should also follow the PCI/DSS guidelines mentioned in the previous section.

The level of activity in the e-commerce fraud investigation process depends on whether the PSPs authorize the payments themselves, or only act as a routing service between the merchants and the original credit card issuers. In the former case the PSPs are more actively involved. In that situation they also hold more of the valuable information to analyse an e-commerce fraud incident. In the latter case they will still be required to connect the payment-related request information from a merchant with the corresponding authorization result coming from an issuer.

If the PSPs hold sensitive information in their databases, they will also be a source of possible data leaks. In that situation they have to put the same precautions in place as issuers have to do (as explained in the next section).

### 3.2.4 Issuer

The issuers are the only members in the e-commerce scenario that know the owners of credit cards in person. Each individual has to register personally with an issuer to get access to a credit card. This registration process includes providing the following information:

- personal related information such as given name, family name and date of birth,

- the currently registered home address,
- the bank account that should be used to settle credit card balances.

Even if the two parties do not really meet each other personally, individuals will still have to identify themselves with a valid ID card and bank account to receive and activate a new credit card. Beside being the single source of truth about the original credit card owner, the issuers of credit cards also collect and store all of their usages. Whereas the Payment Service Providers can only provide individual credit card usage patterns for the online shopping scenario, the issuers can also include those transactions that the credit card owners do in the real-world. Needless to say that these are valuable information for an e-commerce fraud investigation.

Still an issuer does not know any details of the transactions that have been made with a credit card yet. As shown in the Section 3.2.3 the issuers receive only an identifier of the merchants a credit card has been used with. Based on public available information about merchants in a commercial register the issuers could come up with at least the retail branch each merchant operates in.

Being the single source of truth about all issued credit cards, their owners and usage patterns make the issuers another high-risk candidates for possible data leaks. They should as well follow the guidelines from the PCI/DSS standards, should incorporate security standards for their IT systems and the processes of operating them, as well as monitor their IT systems actively with an intrusion detection mechanism.

### 3.2.5 Acquirer

The acquirers hold the bank accounts of merchants and are responsible for withdrawing the outstanding amounts of transactions from the accounts of the consumers, or more precisely requesting it from the issuer of each consumer. Due to this an acquirer usually does not process any credit card related information from consumers directly, but refers to the unique payment tokens that have been created by the PSPs or the issuers during the authorization processes.

Still as financial institutions acquirers (like issuers) have to comply with the rules and guidelines of the PCI/DSS and other industry standards to make sure that their bank accounts as well as the transaction processing are safe and secure. The detailed analysis of these techniques and procedures as well as possible banking fraud incidents are out of scope of this Master Thesis though.

### 3.2.6 Logistic Service Provider

The Logistic Service Providers have two important roles in the e-commerce scenario. First, they have access to and control over the items of a merchant for the duration of the transport between the merchant's facility and the consumer's shipping address. And second, they hold the information to whom they have handed over the items at the final destination. Although the LSPs have nothing to do with any payment-related activities, they are still critical actors for the investigation of fraud incidents, because they will be the last chance for a merchant to stop the delivery of an order (in case a fraud has been detected after initiating the shipment), or provide information about the person that has received the items at the shipping address — especially so for orders of high-priced goods, which usually require a recipient to identify with a personal ID card and place a signature on the delivery receipt.

For initiating the shipment procedure a merchant orders a certain transport service from a LSP, and hands over the following information:

- name of the recipient,
- delivery address given by the consumer,
- list of items to be shipped,
- optionally: value of the items if an insurance policy is taken.

The LSP returns a unique tracking number for the shipment in response. This number can be used by the merchant, and the consumer, to check for the status of a shipment online.

As the LSPs do not have to deal with the payment-related activities in the e-commerce scenario, they are also not actively involved in the fraud investigation. However, they can stop the delivery of the items, or provide useful information about the recipients if an incident is found.

### 3.2.7 Cloud Service Provider

The Cloud Service Providers offer IT services to their customers. These IT services include hardware and software assets that merchants can order in the e-commerce scenario to run their Web shops on the Internet. Part of the service level agreement between a merchant and a CSP is a detailed listing of the responsibilities of both parties (who has to take care of what). In most cases the merchants are outsourcing

the complete operation of the hardware and software for their Web shops to the CSPs; so the CSPs are responsible for making sure that the Web shops are available and secure. The CSPs are also constantly monitoring the incoming connections to each public Internet server under their control and can provide information whether a Web shop of one of the merchants has been compromised or not. However, the CSPs are not actively involved in the e-commerce fraud investigation.

### 3.2.8 Independent Software Vendor

The Independent Software Vendors design, implement and sell the Web shop software tools. They have detailed knowledge about the software components and libraries used within their Web shop products and check them regularly for security breaches or vulnerabilities. They also have to verify these software parts for vulnerabilities that they have implemented on their own, as well as have to make sure that their implementations follow industry standards (e.g. PCI/DSS for handling person and payment-related information). Therefore, they can best assert these quality criteria of a Web shop software if needed. Thus, the ISVs are not an active member of an e-commerce fraud investigation.

### 3.2.9 Internet Service Provider

The Internet Service Providers offer services to the consumers so that they are able to connect to and make use of the Internet. Each Web request consumers are doing on their systems is routed to the Internet via the infrastructure of an ISP. Due to existing regulations and laws the ISPs have to store the log files of each Internet session of their customers for a certain amount of time. Especially, these log files can be helpful to decide whether a consumer was visiting pages in the dark-side of the Web, or if they fall victim to some phishing attacks (explained later in Section 3.4). Although these information can be helpful to decide on fraudulent transactions in the e-commerce scenario, the ISPs are not actively involved in the investigation of it. They are rather required for getting information about the deceivers in case a fraud has been detected.

## 3.3 Data flow for credit card transactions

As explained in the previous section, there are a many stakeholders involved in providing IT hardware, software and services to keep the Web shops on the Internet up and running. Only a small fraction of them will have to deal with the handling of credit card payments and order fulfilments though. These are the relevant stakeholders to

look at in the case of an e-commerce fraud incident. The actual flow of information between them is displayed in Figure 3.3.

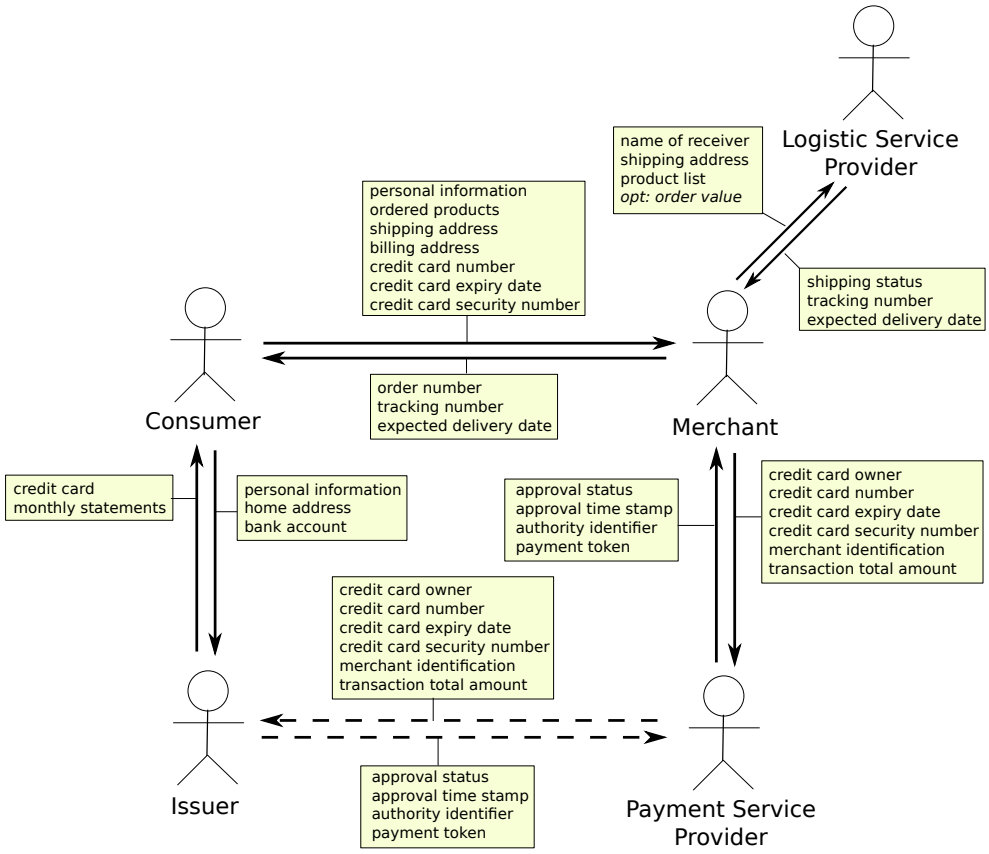


Figure 3.3: Stakeholders and data flow in e-commerce scenario

## 3.4 E-commerce fraud incidents

Based on the explanations in the previous section one can come up with the strategies fraudsters might use to trick the e-commerce system. To do so criminals will have to get access to credit card information first. Therefore, this section starts looking into ways a criminal might get access to credit card and personal related information in the e-commerce scenario. After that the section describes possible strategies fraudsters can use to cheat the system. The section ends with a discussion of the e-commerce fraud incident handling as it is in place today.

### 3.4.1 Credit card data breaches

Following the description in Section 3.3 one can figure out the participants that have access to or store credit card information in the e-commerce scenario, namely:

- a consumer as owner of a credit card,
- an issuer, who handed out a credit card to a consumer,
- a merchant in case a consumer is paying with credit card,
- a Payment Service Provider when a consumer is paying with a credit card online.

The PSPs receive credit card information from merchants with the payment authorization requests. If the PSPs do the authorization themselves, they are also the participants that store and hold the credit card information in their databases. As mentioned earlier the PSPs should follow industry standards and guidelines for storing and processing payment-related information, especially the PCI/DSS standard (Virtue 2009). Additionally, they are responsible for monitoring their systems with an intrusion detection program. This utility will trigger a signal as soon as an hacker got access to the internal databases. In that case the PSPs can put the leaked credit card information on an internal blacklist, so that these cards can no longer be used for further payments online. Moreover, they will have to send a message to the corresponding issuers. The PSPs generally maintain a strong business relationship to each of them. Additionally, the issuers will inform the affected credit card owners, and send out a new credit card to each of them. Due to this procedure in place one can assume that the safety and security of credit card handling at the PSPs can be guaranteed.

The merchants receive the credit card information during a checkout process from a consumer. As the credit card information is transferred via the public Internet from consumers to merchants, it could be victim to man-in-the-middle attacks, in which hackers are intercepting the communication between both stakeholders with the objective to capture personal and payment-related information from the data transmission streams. Therefore, the merchants should offer their Web shops via a secure communication channel only. To do so, they can use industry standards such as TLS to encrypt the information that is send between both parties. Doing so will make it more difficult for attackers to get to the plaintext information exchanged between consumers and merchants during the checkout process. As the merchants are not processing the credit card information directly, they also do not have to store them in their own databases. The merchants ask the PSPs or the issuers of the credit cards for authorization of a payment and receive a unique payment token in response, if that authorization was successful. As stated in the PCI/DSS standard (Virtue 2009) merchants should *never* store credit card information as a whole in their own databases, but should use the unique payment tokens and shortened credit card data (especially truncate credit card

numbers to the last four digits) to refer to a specific payment later. Due to this procedure in place one can conclude that breaking into the systems of a merchant will not result in any leaked credit card information, if merchants follow these guidelines.

The issuers are a valuable target for hacking into their IT systems with the objective to leak a massive amount of credit card and personal related information too. As financial institutions the issuers also have to follow a huge set of regulations and safety procedures to be able to participate on the market. It can be assumed that at least the same safety mechanisms are valid as are in place for the PSPs. This means constantly monitoring the internal systems with an intrusion detection mechanism and blacklisting any leaked credit card. In addition to the monitoring of all online activities (as also the PSPs do) the issuers can monitor activities done with the credit card in the offline world too. In case of suspicious activities the credit cards can be blocked immediately, and new cards can be send out to each affected owner.

The consumers are also a valuable target for eavesdropping on credit card and personal related information. They are also the weakest and most insecure party in the whole e-commerce scenario. As shown before, a lot of the protection mechanisms of the other participants rely on following industry standards, and on constantly monitoring the own systems for malicious activities. This can not be securely said about the computers of the consumers though. Whether they are using up-to-date security programs (e.g. an Anti-virus tool and a firewall) on their computers or not, is out of reach of the other actors to verify. Additionally, consumers can fall victim to phishing attacks, which will send them to malicious Web sites with the objective to get their personal related information. In some seldom cases the consumers might cooperate with fraudsters, or might be the impostors themselves, with the intent to trick the system for their self-interests. Due to these circumstances an e-commerce fraud investigation can not rely on information from the consumers at all, but instead has to figure out if a suspicious transaction was triggered from the owner of the credit card, or whether the transaction was coming from a deceiver.

### 3.4.2 E-commerce fraud strategies

After fraudsters have got access to leaked credit card information, they can come up with the following strategies to trick the e-commerce system:

1. a deceiver owns information about **one** leaked credit card and try to use it for ordering products from **multiple** merchants on the Internet,

2. a deceiver owns information about **multiple** leaked credit cards and try to use them for ordering products from **one** merchant on the Internet,

3. a combination of the two cases above, which can also be related to a series of the first fraud activity.

In the first scenario, in which the fraudsters try out a leaked credit card for ordering products on Web shops of various merchants, each of the merchants only sees the transaction that takes place in their systems. This will make it more difficult for each merchant to detect whether this is a fraudulent transaction or not, because they are not aware of the attempts the fraudsters did on other merchant's Web shop.

As each merchant will rely on a PSP or an issuer to verify the credit card payment, it is in the responsibility of these participants to recognize fraudulent transactions in this specific scenario. To be able to do so, the PSPs and the issuers are monitoring the usage of credit cards and are actively looking for suspicious activities. The fraud prevention mechanisms in place are mostly working on rule-based, and in some cases also on score-based systems running in the internal networks of the PSPs and the issuers. These systems are fed with the information the merchants send with the payment authorization requests and will come up with a decision on each transaction that is either:

1. Yes, this looks like a fraudulent transaction and has to be blocked.

2. No, this seems to be a valid transaction and should be acknowledged.

3. Maybe, this transaction might be valid, but there is some uncertainty in the validation of it. These edge cases are routed to a human operator of a PSP or an issuer to decide how to proceed with them.

As a recent study shows, the success rate of the fraud prevention systems heavily relies on the techniques used to validate the transaction data (Rana & Baria 2015). The outcome is that about 70 to 80% of the fraudulent transactions will be currently recognized as such and blocked successfully. That still means up to 30 percent of fraudulent transactions could not be identified correctly. For investigating these edge cases, each organization employs special trained staff, which is operating 24/7 and 365 days a year, for handling them.

As stated in the introductory of this Master Thesis in Chapter 1, there is a shift from the offline credit card fraud to the online world. This is also resembled in current

figures of e-commerce fraud incidents, which show that it makes up to 85 percent of all credit card fraud attempts and have on average a transaction value of 500 to 600 EURO.

As the PSPs and the issuers do not have any order details, they can only decide on the information given during the payment authorization requests (see Section 3.2). At most they can validate the branch a merchant is operating in, and it might come as no surprise that the fraudsters are regularly using Web shops of merchants that offer either electronics, clothings, entertainment, or travel-related products and services. These are also the most commonly used sources of *valid* e-commerce transactions, and will therefore make any fraudulent transaction very difficult to detect.

At the end it might be an owner of a credit card, who detects suspicious activities on the credit card account and informs the issuer about them. Based on current regulations and laws the issuer has to rollback the fraudulent transactions on request of the consumer, which means that the merchants will have to cover the costs of the e-commerce frauds (as they are not receiving the money for the products that might have been shipped to the fraudsters already).

Looking at the second scenario of the e-commerce fraud strategies at the beginning of this section, a merchant will receive multiple requests from a deceiver, who is trying out various leaked credit cards for finishing an order. These kind of e-commerce frauds can be recognized at the systems of the merchants based on the same source IP address of the requests, or due to having the same shipping address for orders with different credit cards. Therefore, one can conclude that also merchants must take an active role in the fraud prevention processes (if they do not do so already), and try to minimize the amount of fraudulent transactions taken place in their Web shops. As this scenario can be likely managed with additional fraud prevention mechanisms at the merchants, and does not need to involve other parties of the e-commerce scenario to figure out the validity of the transactions, this second scenario falls out of scope of this Master Thesis.

### 3.4.3 Handling of e-commerce fraud incidents

If the fraud prevention systems at the PSPs or the issuers detect a suspicious transaction, an operator working in a special department within the organization will be informed about that transaction via a notification on his or her computer. This operator will have to decide, whether the transaction looks valid and should be acknowledged, or seems to be fraudulent and has to be denied. To be able to decide this, he

or she is going to look into the recent usages of the credit card in question. Whereas it will be easy to recognize that a credit card, which was just being used in a shop in Germany, could not be used in a shop in US or Asia within a short time-frame due to physical constraints in the real world, the same consumer can order products from an US or Asian online retailer with ease within minutes. So these initial geographical constraints, which work so well with real-world usage patterns of credit cards (a proven fraud prevention mechanism called geo-fencing), will no longer work in the e-commerce scenario. Also blocking incoming authorization request based on the geographical location, which is related to the IP address of the consumer, will not be sufficient as deceivers are able to fake them with ease.

So the operators have to found their decisions on the transaction information at hand. Initially, they can check for the amount that has been paid with the credit card. One can assume that small amounts will be covered by the PSPs or the issuers, which will take over the risk for a false payment authorization. But with an increased value of the items ordered, the PSPs and the issuers are putting back the risks to the merchants in case of any customer complaints later. At a second glance, the operators can also verify whether a customer has had any business relationship with a merchant in the past or not, as well as check for the retail branch a merchant operates in. But these are weak hints for investigating the validity of an e-commerce transaction, because they can be bypassed by the fraudsters with ease (see the explanations in the previous section).

To make a solid decision, the operators will have to get in contact with all the merchants a credit card has been used with recently, and have to ask for additional information such as:

- Does the consumer owns an user account with the merchant's Web shop?
- What is the consumer usually looking for in the merchant's Web shop?
- Does the shipping address match the billing address for that order?
- If not, has the consumer send orders to this shipping address in the past?
- What has been ordered by the consumer, incl. detailed product information such as brand, model, product categories, ...?

In some cases the PSPs or the issuers have had a business relationship with an online merchant in the past. So the operators from the PSPs or the issuers might already

know whom to contact from the support personnel of that merchant. But in most cases the contact persons might not be known to them, so they have to send a request to the general support staff via the contact forms on the merchant's Web site.

Getting the right information will still take time, because the correct addressees from the support departments of the merchants are unknown, the merchants do not have specialized staff at hand to handle these kind of inquiries, or there might be misunderstandings on handling a request due to language barriers or different incentives between the participants. Additionally, the operator responsible for the incident has to collect all available information from these merchants, notes them down, and tries to build a "big picture" out of them. In case the initial information received from one of the participants has not been enough, the operator has to get in contact with the support personnel again. This can result in a lengthy sequence of communication attempts and question-response processes between an operator and the online merchants concerned. Thus, getting an in-depth overview of suspicious credit card usages in the e-commerce scenario is likely taking hours, if not days or weeks. That is definitely way to much time and effort to look into any of these suspicious transactions in detail. Therefore, one can assume that an in-depth analysis of any suspicious transaction will not take place today; most of these transactions will be acknowledged without any doubt after a first short look and plausibility check by the operators instead.

However, the merchants as well as the PSPs and the issuers have a high incentive for increasing the success rates of their fraud prevention mechanisms, and for keeping the numbers of successful fraudulent activities low. For the PSPs and the issuers there are regulations stating that at maximum only one thousands of the overall transactions[1] can be fraudulent. This keeps the pressure on these financial institutions to invest in fraud prevention techniques for being able to stay in business. Additionally, it is of high interest for the merchants to resolve a fraudulent transaction before a deceiver receives the ordered products. In the worst case scenario just *one* successfully performed fraudulent transaction in an e-commerce shop will trigger hundreds if not thousands of subsequent attempts from other fraudsters, as past experiences have shown.

## 3.5 Scope of this Master Thesis

As laid out in the previous section, the most interesting e-commerce fraud scenario is the one in which fraudsters use leaked credit card information to order products or services from various merchants on the Internet. This is currently most likely to be

[1] Note: numbers stated are valid for the EU.

successful, because there is a lack of information on the side of the merchants as well as the PSPs and the issuers. Each of the affected merchants just noticed the transaction that takes place in their own Web shop, and are not aware of the other attempts the fraudsters do on the Internet. The PSPs and the issuers will both notice the active use of a credit card on different Web shops though, but do not have any transaction details. Therefore, they could not correlate the data from these transactions to check for suspicious activities.

Based on the current credit card usage patterns of the fraudsters, who will try a leaked credit card to order products from commonly used Web shops, it is more likely that these fraudulent transactions will not be recognized on time by the existing fraud prevention techniques, which work on predefined rules or computed probabilities.

A simple approach to solve these issues would be to just share more information of the ongoing transactions between the merchants, the PSPs and the issuers. This approach might be subject to fail though, because adapting and harmonizing the communication interfaces between the Web shops from various online merchants and the Web Service interfaces of different PSPs and issuers are an enormous undertaking. Any attempt to do so will likely not succeed, because of different notions of the communication patterns and data structures that have to be exchanged between all relevant participants.

To solve these problems, this Master Thesis will look into the information sharing issues in detail and try to come up with a solution to answer the most important question of this scenario:

*Is this transaction really a valid e-commerce transaction?*

Looking into the stakeholders, who can provide useful information to decide it, one will come up with:

1. **Merchants**, who can provide additional information of each e-commerce transaction in question.
2. **PSPs/issuers**, who have information about the credit card usage patterns and the original credit card owners.
3. **LSPs**, who can offer information about whether an order has already been shipped or not, and in the former case to whom it has been handed over.

Its important to point out that parts of the shared information are confidential or business-critical to at least one of the stakeholders involved. Due to this fact, the data

sharing has to be secured, and access to the resources has to be granted to selected participants of the scenario only. This Master Thesis will focus on the data sharing, collecting and combining aspects of the collaborative system. A detailed discussion of the security aspects of it, incl. how to restrict access to the data with available techniques such as OAuth, is out of scope of the Master Thesis though.

Additionally, as a CSCW system *always* consists of a social and a technological component, introducing such a collaborative system into an existing organization will raise issues of user acceptance, and the adaptation to business processes. Due to time and space constraints though, these aspects are also left out of the discussions in this Master Thesis.

# 4 Theoretical Foundations

This chapter lays out the theoretical foundations for the collaborative system that will be designed in Chapter 5 and Chapter 6. It starts with an investigation of the CSCW system theory followed by a detailed examination of the Semantic Web standards such as RDF, OWL and SPARQL. Last but not least, the chapter looks into core concepts of P2P communication technologies and protocols such as WebRTC.

## 4.1 Computer-Supported Cooperative Work

This section gives a short introduction to the theoretical foundations of CSCW systems. It starts with an overview of the research field itself followed by a description of the types of CSCW systems available. After that, it introduces the concept of shared information spaces.

### 4.1.1 Fundamental aspects

CSCW is "a *generic* term which combines the understanding of the way people work in groups with the enabling technologies of computer networking, and associated hardware, software, services and techniques" (Borghoff & Schlichter 2000, pg. 92). It is part of the research field of *Cooperation Systems*, which emerged during the early 1980s with the understanding that a multi-disciplinary approach for designing IT systems is needed for the success of such systems. Thus, the research field looks into the usage of applications to support group work in an organizational setting, the effects of such a system on individual users, as well as how the applications have to be adapted for the group's context. Therefore, the studies of Cooperation Systems consist of a *social* part as well as a *technical* part, and look into the interrelationship between them for certain aspects of work in general and explicitly for communication and cooperation in a team (Grudin 1994).

These systems are generally focused on the concept of human-centred computing that wants to establish technology and work methods to improve processes and results of the work, while also improving the human conditions at work. A "work system" in

such a sense describes the process of human work that consists of goal-directed activities in a *professional* context. As more work has been moved to information workers, another important aspect of such a system is the human cognition, which results in a need of taking human behaviour as well as individual goals and knowledge into consideration. Therefore "work systems" focus on activities of a group of people, a team, an organization or a society, and include social factors such as knowledge, goals, tasks and the work of individuals or subgroups. These "work systems" are getting more and more complex as the problems human beings have to deal with are getting more difficult; in terms of their dynamic, non-linear, interactive and simultaneous nature. Therefore, humans continually have to adapt to their environment, take over different roles, and are engaged in various activities, which include the management of the technologies used, and to handle the issues they introduce (Hoffman et al. 2009).

That said a "work system" in this sense has two possible outcomes: the products and services created together by humans and/or machines, and the sociological and psychological consequences as a result of being part of the process. The objective of a *sociotechnical system* is to optimize both outcomes (see Figure 4.1).

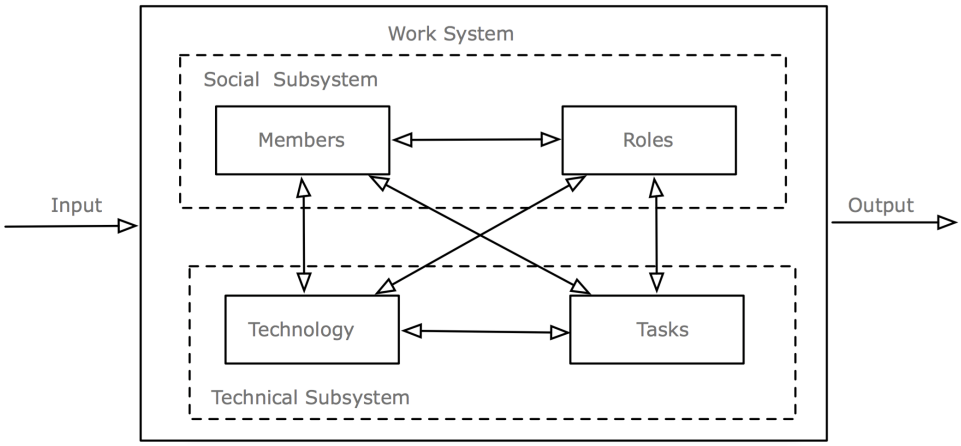


Figure 4.1: A sociotechnical work system (Sydow 1985, pg. 29)

To sum up *sociotechnical* refers to the interrelatedness of social and technical aspects of an organization. The sociotechnical theory is founded on two main principles (Koch 2008):

- The interaction of social and technical factors creates the conditions for successful organizational performance. This interaction consists partly of linear "cause

and effect" relations and partly from "non-linear", often unpredictable relations. Whether designed or not, both types of interaction occur when socio and technical elements are put to work.

- An optimization of each aspect alone (socio or technical) tends to increase not only the quantity of unpredictable relations, but those relations are injurious to the system's performance.

Due to these facts, the success of a *collaborative system* depends highly on the level of active use of the system by its users. To improve the situation the system has to offer a clear balance between efforts and benefits for all of its users, has to communicate those clearly to them, and requires a human-centred user interface as well as a good integration into the context of its users (Koch 2008).

### 4.1.2 Classification of CSCW systems

In general CSCW system can be classified based on the type of communication they support, the kind of applications they make possible, as well as according to the "3C model".

In a distributed team environment the style of communication could be *synchronous* or *asynchronous* depending on the dimension of time. If the communication takes place at the same time the communication is synchronous, otherwise asynchronous. Another aspect that needs to be taken into account is the place. The team can be either co-located or geographically dispersed, which has a huge impact on the type of communication suitable. Taking both dimensions into account leads to the quadrant shown in Figure 4.2.

Moreover, it is possible to group the CSCW systems based on the "3C model" as visualized in Figure 4.3 into (Borghoff & Schlichter 2000, pg. 125):

- **communication support:** for a two way exchange of information between different team members,
- **coordination support:** for management of shared resources such as meeting rooms, network printers, file shares, ...,
- **collaboration support:** to enable members of a group work in a shared environment to reach a predefined goal.

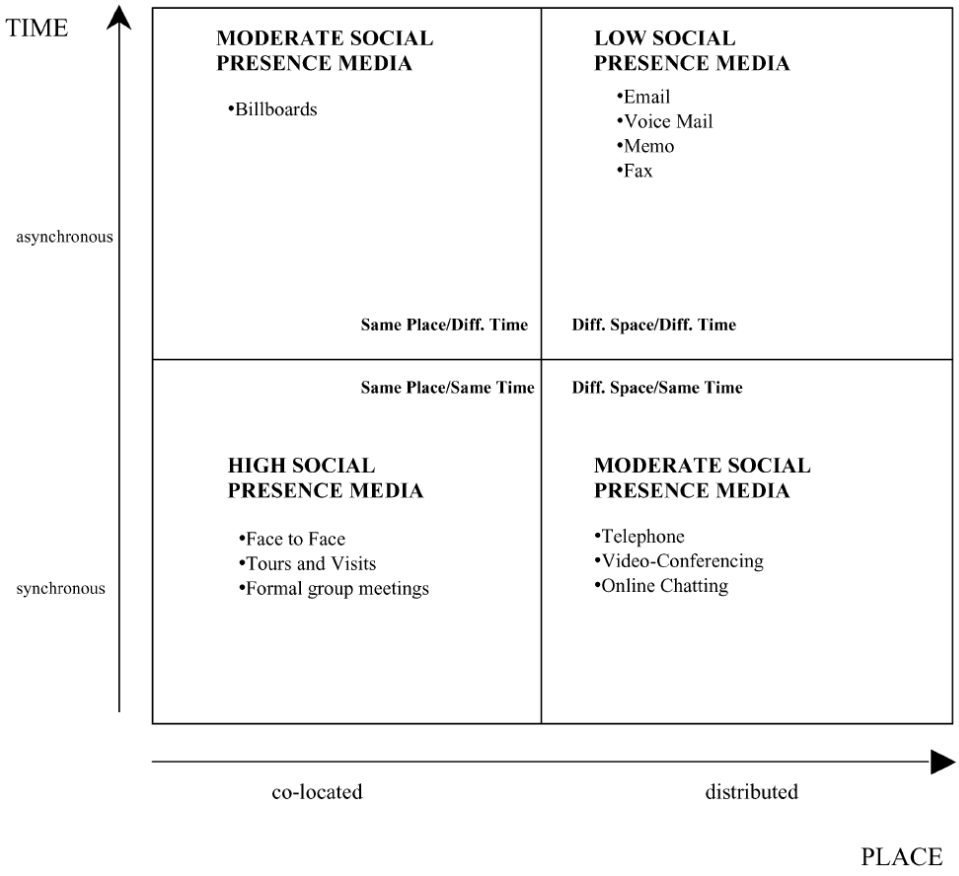


Figure 4.2: Time/Place Matrix (Robert & Dennis 2005)

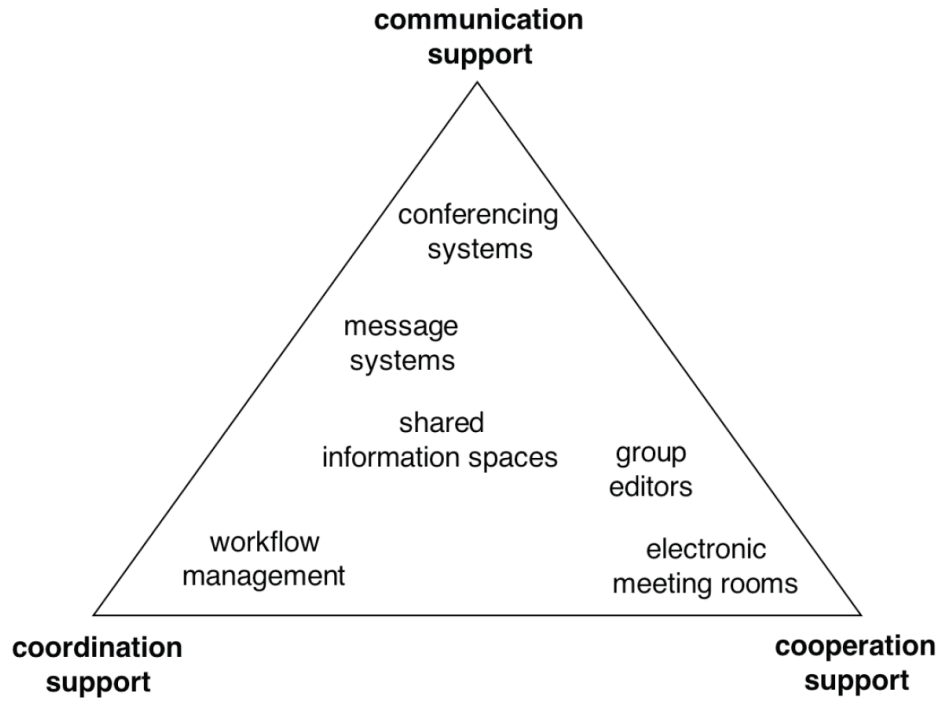


Figure 4.3: The 3C model (Koch 2008)

Typical application classes for CSCW systems might be (Borghoff & Schlichter 2000, pg. 119-120):

- **message systems:** allow to exchange textual messages between team members asynchronously; modern systems allow sending of other digital artefacts such as images and documents as well (e.g. instant messengers such as Microsoft Skype or Slack),
- **group editors:** allow collaborative work on some kind of shared document or artefact; editing of the shared document can be either allowed synchronously at the same time, or also asynchronously at different times (e.g. collaborative word processors such as Microsoft Word Online or Google Docs),
- **electronic meeting rooms:** allow multiple participants to work within a shared workspace or on a shared whiteboard, and offer support for ad-hoc brainstorming, idea generation and group decision making,
- **shared information spaces:** allow participants to access information at any place any time as well as to share information with others (e.g. Microsoft Sharepoint)

### 4.1.3 Shared Information Spaces

In a CSCW system shared information can play two important roles: on the one hand they can transfer knowledge and facts between participants, and on the other hand they resemble intermediate or final results of the group works themselves. In case of business critical information the system also has to provide a log of activities as well as a history of all artefacts generated or manipulated with it over time (Borghoff & Schlichter 2000, pg. 295).

The kind of artefacts that are shared within such as collaborative system is not further specified and can range from information, events or object representations from the real world to internal terms or objects of the working group. Whereas the former can be described extensively and usually does not need further interpretation, the latter needs an interpretative component to define and communicate their intended meanings between group members. This common interpretation of terms and objects is even more important, if the collaborative system works across time and space boundaries, because co-located group member generally have the same understanding of terms and objects due to being in the same (working) context and environment. Based to this fact, information that has to be shared between dispersed group members has to be

refined and packaged in a way that enable the receiver to “unpack” the information and to re-create the original context, in which the information was created. Therefore, the information in such a system is not just coming out of a shared database, but also implies the joined interpretation of it by all the actors involved (Bannon & Bødker 1997).

If information from different sources comes together in a shared information space, the collaborative system should support a non-linear, exploratory way to retrieve and navigate through the information space to enable participants to browse and ascertain the concepts and their relations individually. A valid proposal for such a system is based on *Hypertext*, because of its generic approach for the construction of non-linear, computer-supported material that users can display and navigate on their screen in a non-linear fashion. Hypertext systems provide information that is distributed over a network of nodes, which make up the information space. Therefore, the information is divided into small, logical information units (aka nodes), in which references (aka links) point to relevant or related units from the shared information space (see Figure 4.4). Users can navigate the information space along these links (Borghoff & Schlichter 2000, pg. 295-307).

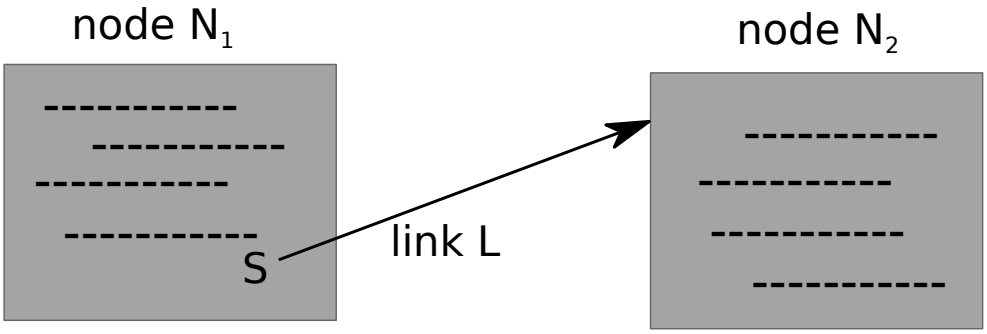


Figure 4.4: A link between two nodes (Borghoff & Schlichter 2000, pg. 303)

A Hypertext system usually consists of three layers (Borghoff & Schlichter 2000, pg. 301-302):

- **Database layer:** persists and manages the Hypertext information in a way that allows fast access and retrieval of information units,
- **Hypertext abstract machine:** creates an information network based on the information units and their relations (aka links between them),
- **Presentation:** displays individual information units on screen and allows the user to navigate the information space (via the links).

The link specifications can be either part of the node content itself (e.g. as in the HTML standard), can be completely separated from the contents of the nodes (e.g. as navigational elements such as "previous" or "next"), or can be collected and presented in some kind of general overview such as a table of contents (Borghoff & Schlichter 2000, pg. 304-306).

## 4.2 The Semantic Web

The *Semantic Web* initiative strives for a better integration of distributed data from various publishers on the Web with the objective to enable new kinds of Smart Web applications. To achieve this goal, the Semantic Web delivers the infrastructure for this vision in the form of various standard specifications such as RDF, RDFS, OWL, SPARQL, ..., which are introduced during the course of this section. Before going into the technical specifications of each of them, the section shows the fundamental aspects underlying the (Semantic) Web as a whole.

### 4.2.1 Fundamental aspects

The Semantic Web builds on the fundamentals of the existing World-Wide Web, especially on (Allemang & Hendler 2011, pg. 4-11):

- **AAA-Slogan:** "Anyone can say Anything about Any topic". The Web does not restrict or control what people post or publish on it. The readers are responsible for deciding whether they can trust information from a specific source or not.
- **Open World Assumption:** As the amount of information on the Web is limitless, and new information is published every day, one must always assume that there is new information available that one does not know yet. Thus, one can never be sure to have all facts at hand. New information can be published at any time that can generate additional insights to the topic.
- **Non-unique Naming Assumption:** There is no central authority responsible for providing unique identifiers for entities on the Web. Due to this fact, different URIs might refer to the same virtual entity or real-world object.

Instead of making information on the Web available for human consumption *only*, the Semantic Web aims to make the information on the Web accessible (and readable) to machines as well. This will allow for the integration of information across Web sites, and enable a distributed, interlinked "Web of Data". The major design principles to achieve this objective are (Antoniou & Van Harmelen 2012, pg. 1-22):

1. to make structured and semi-structured data available in standard formats,
2. to make individual data elements and their relations accessible on the Web,
3. to describe the intended semantics of the data in a machine readable format.

The data model of the Semantic Web is build upon labelled graphs with objects and their relations. Objects are modelled *nodes* and their relations as *edges* between them. To express these graphs of related objects, the Semantic Web has to (Antoniou & Van Harmelen 2012, pg. 1-22):

- formalize the syntax of the graph in RDF (see Section 4.2.2),
- use URIs to identify individual data items and relations,
- use ontologies to represent semantics of the entities. Ontologies can be lightweight RDFS definitions or expressive descriptions in the OWL language (see Section 4.2.3).

Initially, one tried to solve the data integration aspect on the Web with the exchange of XML-based messages, but though the XML format is more machine readable than HTML, it still lacks the semantic of the data transmitted. Thus, the Semantic Web defines the RDF format as its basic data exchange format. Still the RDF format was initially based on the XML specification. To formally describe the existing terms and their possible relationships within a domain the Semantic Web relies on an ontology specification. These specifications are either expressed in RDFS, or by using the more expressive OWL language; both are meta-description languages, which allow the definition of domain-specific knowledge representations based on the concepts found in RDF itself.

Due to this, the Semantic Web is a layered approach as depicted in Figure 4.5.

### 4.2.2 A Resource Description Framework

When trying to come up with a specification on how to integrate data on a globally dispersed platform such as the World-Wide Web, one will have to answer the following questions (Antoniou & Van Harmelen 2012, pg.23-25):

- **Syntax:** How to serialize the data?
- **Data model:** How to structure and organize the data?
- **Semantics:** How to interpret the data?

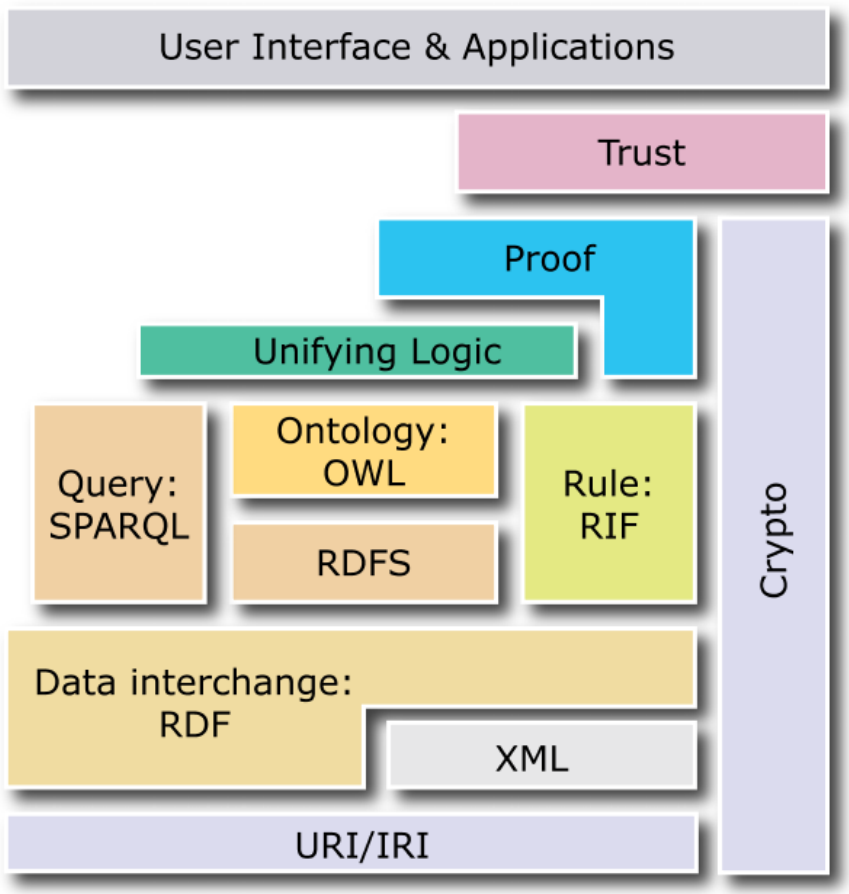


Figure 4.5: The Semantic Web model (W3C 2013)

Whereas the World-Wide Web is made up from interlinked documents in the HTML format that are specifically designed for rendering information on screen, and will be consumed by a human, the RDF brings a highly flexible data model to the Web. Its basic building block is the *triple*, which is a statement consisting of an entity, an attribute, and a value. The individual parts of a statement are also known as subject, predicate, and object, which make up a directed graph as shown in the example in Figure 4.6:

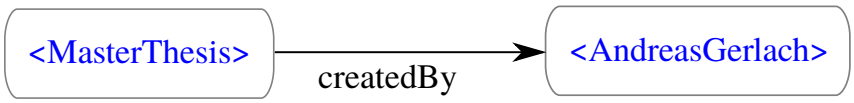


Figure 4.6: A basic example for a triple-statement

In this example the triple consists of the entity “MasterThesis”, the assigned attribute “createdBy” and the value “AndreasGerlach”. The value-part of a triple can contain either a literal value or refer to another entity (as in the example depicted above). However, this example statement is problematic, because the entities are not unique. Based on the given information it is not clear, which specific “MasterThesis” is meant, and to whom the value “AndreasGerlach” refers. Furthermore, the predicate used can have multiple meanings. These ambiguities have to be resolved on the Semantic Web to be able to make these information understandable to machines. To solve these issues, the Semantic Web standard specifies that names of entities and predicates have to use a URI to make their intended meanings clear. Literals that can also be used as values, such as numbers, dates and strings, borrow their data type specifications from the XML standard (Wood et al. 2014, pg. 15-38).

Based on this description the foundational elements of RDF can be summarized as:

- **Entities:** aka resources or “things of interest” that are identified via URIs,
- **Predicates:** aka attributes or properties that specify the relations between resources and are also identified by URIs,
- **Literals:** integral values such as numbers, dates and strings that are based on the XML data type specification,
- **Statements:** assign a value (either another entity or a literal) to a “entity-predicate” relation,
- **Graphs:** are the data model behind RDF and enable a distributed, interlinked “Web of Data”.

RDF triples can be serialized into four different syntax formats (Wood et al. 2014, pg. 43-54):

- **RDF/XML:** the original format of the RDF data sets based on the XML specification. Because of their complexity they are best used with a parser program. For an example see Listing 1.
- **RDFa:** describes how to embed RDF information into existing HTML documents. It allows authors to enrich their Web pages with semantic information by adding a set of predefined HTML attributes to important items within the document. Listing 2 shows a basic example.

- **JSON-LD:** a recent initiative to allow JavaScript developers to use JSON documents to express a RDF data set; see Listing 3 for an example.
- **Turtle:** a human-readable serialization format for RDF statements. URIs can be shortened with a predeclared prefix, statements have to end with a period. Statements referring to the same entity can be abbreviated via a colon, which repeats the subject from the previous statement, or a comma, which repeats subject and predicate from it. For an example see Listing 4.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:ex="http://www.example.com/">
  <rdf:Description rdf:about="http://www.example.com/MasterThesis">
    <ex:createdBy rdf:resource="http://www.example.com/AndreasGerlach"
  ↪  />
  </rdf:Description>
</rdf:RDF>
```

Listing 1: A triple statement expressed in RDF/XML format

```
<div about="http://www.example.com/MasterThesis">
  <span rel="http://www.example.com/createdBy"
↪  resource="http://www.example.com/AndreasGerlach">
</div>
```

Listing 2: A triple statement expressed in RDFa format

```
{
 "@context": "http://www.example.com/",
 "@id": "http://www.example.com/MasterThesis",
 "createdBy": "http://www.example.com/AndreasGerlach"
}
```

Listing 3: A triple statement expressed in JSON-LD format

```
@prefix ex:  <http://www.example.com/> .
ex:MasterThesis ex:createdBy ex:AndreasGerlach .
```

Listing 4: A triple statement expressed in Turle format

Interestingly, these different serialization formats for RDF data sets are fully interchangeable. A RDF data set can be easily converted from one serialization format to the other, and merging RDF data sets work with sources expressed in different formats as well.

Coming back to the initial questions that have to be solved for data integration on Web scale, the section showed how the Semantic Web initiative tries to solve them, such as this (Antoniou & Van Harmelen 2012, pg.23-25):

- **Syntax:** support for the following formats: Turtle, RDFa, RDF/XML and JSON-LD,
- **Data model:** use the graph-based data model of RDF,
- **Semantics:** express the semantics of the data in RDFS.

This will be the topic of the next section.

### 4.2.3 RDF vocabularies and Web Ontologies

For describing domain specific semantics of the data in a RDF data set one can either use the lightweight RDFS standard to define available entities and their relationships, or use the more expressive OWL specification from the W3C.

Both specifications are based on the RDF data model, and make use of the following predefined URIs to declare their meanings, see Table 4.1:

| **Name** | **Prefix** | **Used for** | **Namespace URI** |
|---|---|---|---|
| RDF | rdf: | Core RDF framework | `http://www.w3.org/1999/02/22-rdf-syntax-ns#` |
| RDFS | rdfs: | Define RDF vocabularies | `http://www.w3.org/2000/01/rdf-schema#` |
| Web Ontology Language | owl: | Define ontologies | `http://www.w3.org/2002/07/owl#` |

Table 4.1: RDF vocabularies specified by the W3C (Wood et al. 2014, pg. 41)

An important step in the definition of a RDF vocabulary or Web ontology is to analyse the domain of interest in detail, and come up with a list of objects and their possible relations. One can use the following step-by-step guide as a reference (Antoniou & Van Harmelen 2012, pg. 40-55):

1. Specify the *things* to talk about. These have to be divided into *objects* (aka real entities) and *classes* (aka set of entities). A specific statement containing the predicate "rdf:type" assigns individual objects to their classes.

2. Look for relations that are available between these classes. The relations can either be of type inheritance or composition.

3. Define existing properties (aka predicates) and their hierarchical relationships (if appropriate).

4. Impose restrictions on the kind of properties that can be used on objects. These can include restrictions on the *values* a predicate might take (aka "rdfs:range" restrictions) and/or restrictions on the possible *subjects* of a predicate (aka "rdfs:domain" restrictions).

The fundamental concepts of the RDFS specification, which are used to model entities and their relations within a domain, are listed in Table 4.2.

| **Classes** | **Used for** |
|---|---|
| rdfs:Resource | individual resources |
| rdfs:Class | classes |
| rdfs:Literal | literals |
| rdfs:Propery | properties |
| **Predicate** | **Describes** |
| rdf:type | kind of class |
| rdfs:subClassOf | inheritance between classes |
| rdfs:subPropertyOf | inheritance between properties |
| rdfs:domain | restrict the subjects of a property |
| rdfs:range | restrict the values of a property |

Table 4.2: RDFS axioms commonly used to define RDF vocabularies (Antoniou & Van Harmelen 2012, pg. 46-49)

As an example Listing 5 describes a RDF data set, which is also displayed in Figure 4.7, as such:

1. there is a class “Song” that is derived from the general class “Audio”,
2. a class of type “Song” can have a predicate named “title”,
3. the predicate “title” is a subproperty of “attribute” and can contain a literal value.

```
# define prefixes for URIs
@prefix ex: <http://www.example.com/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

# define available classes and their hierarchy
ex:Audio rdf:type rdfs:Class .
ex:Song  rdf:type rdfs:Class;
         rdfs:subClassOf ex:Audio .

# define available properties, their hierarchy and restrictions
ex:attribute rdf:type rdfs:Property .
ex:title rdf:type rdfs:Property;
         rdfs:subPropertyOf ex:attribute;
         rdfs:domain ex:Song;
         rdfs:range rdfs:Literal .

# add specific instances of classes
<http://music.org/Song1> rdf:type ex:Song;
                         ex:title "The best of..." .
```

Listing 5: A sample RDF data set based on Figure 4.7

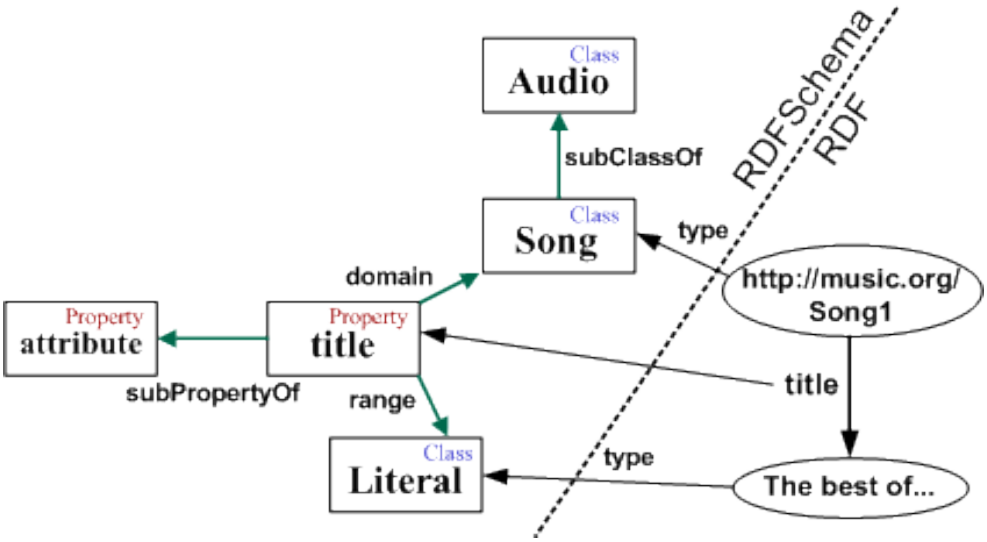


Figure 4.7: RDF Schema sample (García 2006)

As is also displayed in Figure 4.7, the majority of the domain specific knowledge is described in the RDFS part of the RDF data set. It holds all available classes, predicates, restrictions, ..., and is therefore more abstract and expressive. The RDF part containing the specific instances might be rather small and concrete, consisting of only existing resources with their known predicates. Usually each of the real entities from the RDF part refer to its best matching and most descriptive entity from the RDFS part (e.g. the entity `http://music.org/Song1` in the sample is referring to the *Song* class rather than the *Audio* class).

Its important to note that the meaning of the RDFS predicates "rdfs:subClassOf", "rdfs:subPropertyOf", as well as "rdfs:domain" and "rdfs:range" is *not* to restrict or validate the proper usage of them in a RDF statement, but is rather used to *infer* additional statements in a RDF data set based on their usage. So in the example above, one can infer that the resource found at `http://music.org/Song1` is not only a "Song", but also an "Audio" based on the "rdfs:subClassOf" relationship between "Song" and "Audio". This kind of propagation of RDF statements based on the usages of those RDFS predicates fits well to the Open World Assumption of the Semantic Web (Allemang & Hendler 2011, pg. 125-152).

In addition to the axioms shown above, the RDF and RDFS specifications contain further useful entities and predicates, such as listed in Table 4.3:

| **Classes** | **Used for** |
|---|---|
| rdf:Bag | unordered list of entities |
| rdf:Seq | ordered list of entities |
| rdf:Alt | list of alternatives or choices |
| rdf:Container | superclass of all containers |
| **Predicates** | **Describes** |
| rdfs:seeAlso | links to an external resource that contains additional information about it |
| rdfs:isDefinedBy | links to the original definition of a resource |
| rdfs:comment | comments and notes on entities |
| rdfs:label | human friendly label for entities |

Table 4.3: RDF and RDFS supplemental axioms (Antoniou & Van Harmelen 2012, pg. 46-49)

As the two tables (Table 4.2 and Table 4.3) show, the RDFS specification contains only basic axioms to describe an entity and its possible relations, as well as the hierarchy between these entities and the relations. It is the fundamental set of constraints that is needed to start with publishing resources on the Semantic Web. However, it does not include more advanced features that can be of use for combining and reasoning on distributed RDF data sets, such as:

- **Equality/Inequality:** Neither RDF nor RDFS provide a way to specify that two resources or properties coming from different RDF data sets are the same, or are not the same. Though it might be possible to model the aspect of equality with a combination of "rdfs:subClassOf" or "rdfs:subPropertyOf" statements, the results are usually not as desired due to the inferencing nature of those axioms.

- **Cardinality:** Predicates defined in RDFS can be used multiple times in statements referring to the same subject or object. This is not desired in situations that only allow one possible instance of a "subject-predicate-object" relation.

- **Transitivity:** Whereas it is possible to express the hierarchy of classes and predicates in RDFS, it is not feasible to do so for individual instances. This might be useful when building up a hierarchy of people, e.g. a statement such as "S3 has an ancestor S2, who has an ancestor S1" can lead to the inference "S3 has an ancestor S1".

- **Property Restrictions:** To define whether a predicate refers to an object or a literal as value is not possible with RDFS, but may be useful for modelling and editing tools.

Therefore, if more expressiveness is required to model a domain, one can use additional concepts from the Web Ontology Language specification (OWL). It is build on the RDFS specification, but contains additional constraints and classifiers; some of the most commonly used ones are listed in Table 4.4.

| **Classes** | **Used for** |
|---|---|
| owl:Class | all classes in OWL |
| owl:FunctionalProperty | allow only one value |
| owl:InverseFunctionalProperty | allow only one source |
| owl:TransitiveProperty | build chains of relationships |
| owl:ObjectProperty | property can hold a resource as value |
| owl:DataProperty | property can hold a literal as value |
| **Predicates** | **Describes** |
| owl:equivalentClass | equality of classes |
| owl:equivalentProperty | equality of properties |
| owl:sameAs | equality of individual resources |

Table 4.4: Commonly used OWL axioms (Allemang & Hendler 2011, pg. 153-185)

Table 4.4 shows only a small subset of the axioms available in OWL. The Web Ontology Language can be used to express a wide variety of constraints and classifiers on predicates. The whole specification is based on three parts that build on each other (W3C 2004). However, with an increase of expressiveness in the model the complexity of the reasoning and inferencing engine also grows, because most of these axioms are used to express inferences that can be drawn on the RDF data set. This Master Thesis restricts the usage of axioms of the OWL specifications to the ones shown in Table 4.4.

### 4.2.4 SPARQL protocol and query language

A kind of query language is required to be able to access specific information in a RDF data set on the Semantic Web. Additionally, the query language has to support the distributed nature of information on the Semantic Web, as well as be suitable for asking information from the graph-oriented data model of RDF. The W3C proposes the SPARQL protocol and query language as a standard way to access and query for information on the Semantic Web. As the name already implies, the specification con-

tains two parts: a *protocol* and a *query language*.

SPARQL requires the RDF data set in a *triple store*, which is a kind of database containing RDF statements (another name for it might be *graph store*). The RDF data set is usually inserted into the triple store via bulk load operations or single SPARQL update statements, which like in SQL for relational databases can manipulate RDF statements in the data store. A SPARQL query is usually expressed in a Turtle-like syntax and is send to an HTTP endpoint of the triple store by a client application. The result of the query can contain either a single value, a tabular data stream, or a subgraph of the RDF data set depending on the type of query issued. The structure of a Semantic Web application that uses a triple store (aka RDF store) and the SPARQL query engine is depicted in Figure 4.8 (Allemang & Hendler 2011, pg. 51-60):

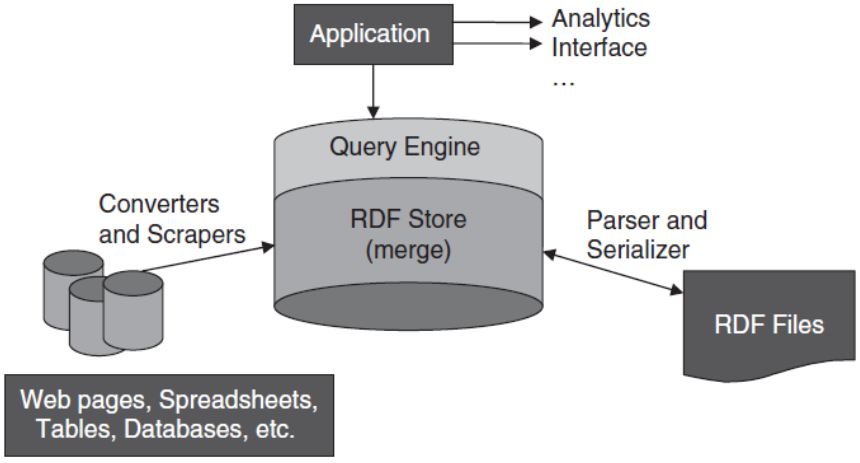


Figure 4.8: Semantic Web application architecture (Allemang & Hendler 2011, pg. 57)

The SPARQL query language has a lot of similarities with the SQL used for querying relational databases. This design decision will ease the transition to the Semantic Web for application developers familiar with accessing data from a relational database. The main difference though is the way used to specify the conditions for a query. This is largely due to the differences in the underlying data model, which is relational in SQL versus graph-oriented in SPARQL. Thus, a WHERE clause in a SPARQL query has to contain a graph-based representation of the query conditions that have to be matched in the RDF data set. Parts of these conditions can be marked as placeholders, which can also be referred to in a SELECT statement for generating a tabular data output. These placeholders are marked with a question-mark in the beginning of their name, and can be used as placeholder for any part of a triple statement (subject, predicate, object) (Allemang & Hendler 2011, pg. 66-112). For querying a RDF data

set containing triples such as the one described in Listing 5 for the titles of all available songs, one has to write a SPARQL SELECT query as shown in Listing 6:

```
# define prefixes for URIs
PREFIX ex:  <http://www.example.com/> .
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

# choose to output any title found in graph-pattern
SELECT ?title
WHERE {
  # describe the conditions for the query
  # as graph-patterns that have to match
  # here: it has to be a Song, which has a predicate title
  ?song rdf:type ex:Song .
  ?song ex:title ?title .
}
```

Listing 6: Selecting the title from all songs with SPARQL

Please note that this query defines two placeholders in the WHERE clause: "?song" and "?title", but only uses the "?title" as an output criteria in the SELECT statement. The placeholder "?song" is *just* required to refer to the same node, when specifying the conditions in the WHERE clause.

The WHERE clause in a SPARQL query can include additional conditions that have an effect on the returned information, such as (Allemang & Hendler 2011, pg. 66-112):

- **LIMIT:** specifies the upper limit of results that should be returned from a SPARQL query; e.g. LIMIT 100
- **FILTER:** express additional filter conditions on the result set of a SPARQL query; e.g. FILTER(?releaseDate >"1980-01-01")
- **UNION:** combine the result sets from different graph patterns into one; e.g. { ex:Song ex:title "A" } UNION { ex:Song ex:title "G" }

Beside the SELECT statement that returns tabular data SPARQL also supports the ASK type of query, which checks for the existence of graph patterns stated in the WHERE clause, and will return a boolean value (true or false). This kind of query is

commonly used to assert the presence of certain triples in a RDF data set.

Another kind of query is the CONSTRUCT statement, which is used to retrieve a subgraph from a RDF data set. It can also be used to harmonise graphs from sources with different schemata. As a query of type CONSTRUCT will return a new RDF graph, it is also used for basic reasoning functionality such as "if this graph-pattern is found, assume that ...", as well as helps to resolve issues of identifying entities that are described with different URIs (Allemang & Hendler 2011, pg. 88-98). For querying a RDF data set containing triples such as the one described in Listing 5 and mapping the song information to the DublinCore specification (Dublin Core Metadata Initiative 2012), one has to write a SPARQL CONSTRUCT query as shown in Listing 7.

```
# define prefixes for URIs
PREFIX ex:  <http://www.example.com/> .
PREFIX dc:  <http://purl.org/dc/elements/1.1/> .
PREFIX dct: <http://purl.org/dc/terms/> .
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

# create a new graph with the mapped song information
CONSTRUCT {
  ?song rdf:type dc:Sound .
  ?song dct:title ?title .
}
WHERE {
  # describe the conditions for the query
  # as graph-patterns that have to match
  # here: it has to be a Song, which has a predicate title
  ?song rdf:type ex:Song .
  ?song ex:title ?title .
}
```

Listing 7: Mapping custom song information to DublinCore vocabulary with SPARQL

Please note that the CONSTRUCT statement consists of a set of triple statements that will make up the resulting RDF graph.

## 4.3 Peer-to-peer communication

This section explains the core concepts of P2P communication technologies. It begins with a comparison of the benefits and disadvantages of centralized and decentralized Web architectures. After that, it shows how P2P communication networks can be structured, the different ways to initiate a communication session, as well as how data can be transmitted between peers.

### 4.3.1 Centralized vs. Decentralized Web architectures

In classical client-server applications the information is stored on a central system (aka server). Clients have to connect to the server and ask for the information. The server handles the requests from the clients and deliver the information in case a request was valid. Prominent examples of centralized Web architectures are Social Networks such as Facebook or Twitter, in which clients such as a Web browser or Mobile application communicate with a Web service, which runs on a server of the organization providing these Social Networks, to access and retrieve Web documents (e.g. HTML, images, audio, video, ...) via the HTTP protocol, as shown in Figure 4.9.

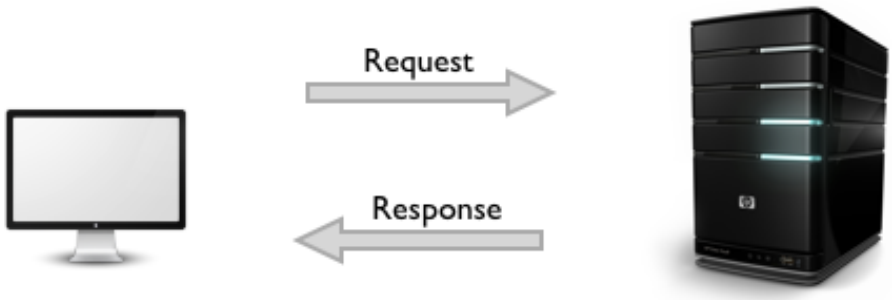


Figure 4.9: Centralized Web architectures as used by prominent Social Networks (Podila 2013)

As a consequence of this architecture all of the information are centralized and under control of the provider of the (Web) service. This can lead to a variety of problems, including serious issues such as unreliable or no longer existing services will result in a dismissal of all the information stored on them, or privacy concerns for user-generated content stored on those central servers.

In opposite to that, a P2P network considers all nodes as equal. This offers the benefits that information can be kept on each node, and each node can provide access to its

information to any other node on the network. Due to this decision, the P2P system has an high degree of decentralization, is not owned and controlled by a specific company, and therefore tends to be more resilient to faults, outages and attacks. But due to the distributed nature of it, looking for and accessing information is more difficult. Information in a P2P system has to be indexed in a way, so that the correct node is queried for it. Moreover, this index has to be stored somewhere in the system, and the optimal solution for the indexing problem depends on the type of P2P system used (see next section). Additionally, the way new nodes get connected to the system is depending on the type of P2P system used, and might lead to the introduction of special bootstrap or super nodes into the network as shown in Figure 4.10 (Parameswaran et al. 2001).

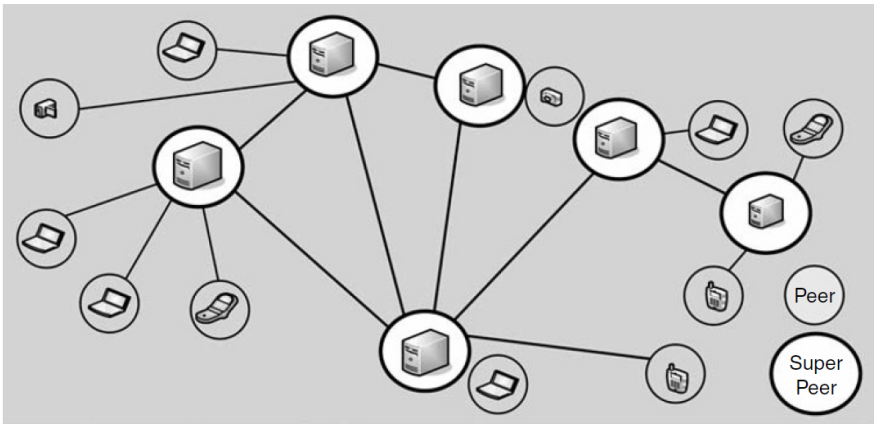


Figure 4.10: A P2P overlay network (Buford et al. 2009, pg. 9)

### 4.3.2 Classification of P2P systems

P2P system architectures can be classified based on their degree of centralization into:

- **Partially centralized P2P system:** rely on a dedicated controller node that maintains the set of participating nodes, host the index of the information available in the system, and controls the overall operation of the network,
- **Decentralized P2P system:** does not use any dedicated controller node, but may need to introduce bootstrap and super nodes for maintaining the list of participating nodes and the index of the information available depending on the size of the network.

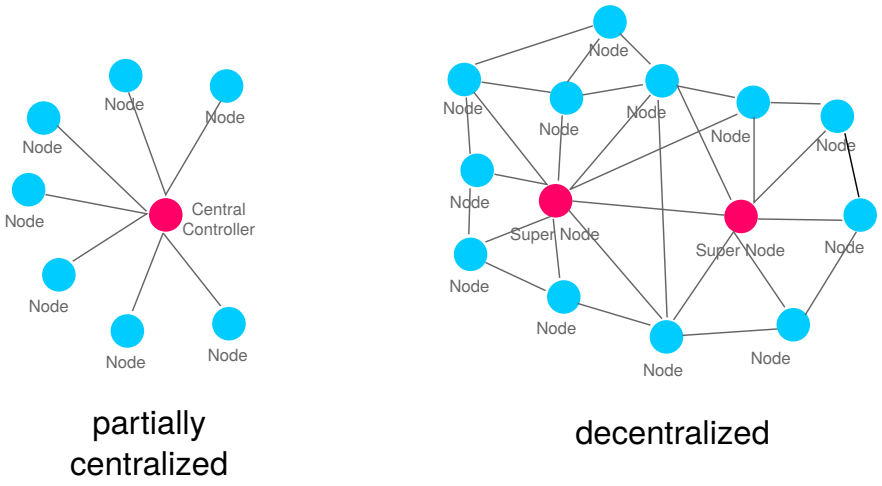


Figure 4.11: Classification of P2P networks

### 4.3.3 Communication in a P2P network

The procedure required to establish a P2P communication depends on the structure of the P2P system. In a *partly centralized P2P system* new nodes join the network by connecting to the central controller first. This central controller has a well-known IP address and maintains the operation of the whole P2P network. Moreover, any new node has to register with the central controller to get introduced to the P2P network. The controller also maintains the information about the overlay network, as well as holds information about each object and on which node(s) it resides within the network. The overlay is typically following a star-shaped topology with the central controller at the centre, see Figure 4.11.

In a *decentralized P2P system* new nodes are expected to obtain the IP address, which they have to connect to initially, via a separate channel (e.g. as a link on a Web site). Depending on the size of the P2P network additional bootstrap or super nodes, which help to setup a new node, are available on the network. These special nodes are generally also consolidating information about the objects available on the peers nearby, which helps speeding up searching and accessing required information. The overlay information of such a distributed network can be either *structured*, in which each node receives a unique identifier from a numeric keyspace resembling the responsibilities of that node, or *unstructured*, in which there is no particular network structure, and no further constraints are assigned to the nodes of the network.

A *structured overlay* maintain the information within the network more efficiently, because it uses a distributed hash-table to maintain a distributed index, and decides

the location (aka node) of an object in the network based on its hash-value. In an *unstructured overlay network* the information is typically stored on the node that introduces it. To locate an object a query request is typically broadcasted through the overlay network. Based on the size of the network and the distance between the node asking for and the node holding the information querying and accessing an information on an unstructured overlay network can take some time, and can also flood the whole network with query requests. Therefore, requesting nodes often set the scope of the request, which limits the number of hops that should be done on the network. This will reduce the communication overhead on the whole system. Additionally, introducing super nodes that collect and maintain indexes of their peers nearby can further reduce the number of hops necessary to find the required information, see Figure 4.11 (Rodrigues & Druschel 2010).

### 4.3.4 The WebRTC standard

"Web Real-Time Communication (WebRTC) is a collection of standards, protocols, and JavaScript APIs, the combination of which enables peer-to-peer audio, video, and data sharing between browsers (peers)" (Grigorik 2013, pg. 307). Although this new W3C standard usually stands for in-browser video or audio conferencing without the need of proprietary browser extensions, it also offers ways to exchange arbitrary messages or binary data between participating peers in a distributed Web application. Due to being an open Web standard, WebRTC is available in many current Web browsers directly, and is widely adopted as a standardized and open way to establish a P2P communication between clients of a Web site, or from within a Web application. The standard wraps a lot of the complexities of establishing peer-to-peer communication channels and transmitting data into three primary APIs (Grigorik 2013, pg. 307-308):

- **MediaStream:** for acquiring access to and retrieve data from local audio and video devices,
- **RTCPeerConnection:** for establishing a peer-to-peer connection between clients,
- **RTCDataChannel:** for transmitting arbitrary application data

To establish a data connection between peers, a Web application has to create a RTCPeerConnection object first, before it can create a RTCDataChannel to exchange messages on it. Establishing a P2P connection between globally dispersed peers on the Web is not a trivial task, and has to provide fallback solutions in case of P2P connectivity issues due to firewall or NAT services used by some of the peers, which

usually prevent clients to connect to each other directly. Fortunately, the W3C standard is taking care of these steps during the initiating of a WebRTC connection by utilizing the ICE protocol. After being able to open a connection to another peer, a communication session has to be created. For that, the communicating peers have to negotiate on protocols, encodings, and additional functionality required for the P2P communication tasks at hand. The WebRTC uses the SCTP to exchange application data between peers (Grigorik 2013, pg. 315-330). It has the following set of features (Grigorik 2013, pg. 342):

- **Reliability:** the data channel can be configured to use either reliable or unreliable delivery of packages,
- **Delivery:** the data channel can be also configured to support either in-order or out-of-order delivery of packages,
- **Transmission:** the transport of data is message-oriented,
- **Confidentiality/Integrity:** all application data transmitted between the peers is encrypted to guarantee confidentiality and integrity of the data exchanged.

For a purely data transmission channel one can also disable any audio and video transfers during the setup of the communication session (see Listing 8).

```
1  // create signaling channel for negotiating between peers
2  var signalingChannel = new SignalingChannel();
3  // create p2p connection object
4  var pc = new RTCPeerConnection(iceConfig);
5  // create a named data channel with unreliable transfer option
6  var dc = pc.createDataChannel('namedChannel', { reliable: false });
7  // set media constraints to disable audio and video transfers
8  var mediaConstraints = {
9    mandatory: {
10     OfferToReceiveAudio: false,
11     OfferToReceiveVideo: false
12   }
13 };
14 pc.createOffer((offer) => { ... }, null, mediaConstraints);
```

Listing 8: Establishing a pure WebRTC data connection (Grigorik 2013, pg. 349)

Once a data channel has been established between the peers, application data can be exchanged between them via message passing, as shown in Listing 9:

```
// initial channel and session setup and negotiation
var pc = new RTCPeerConnection(iceConfig);
var dc = pc.createDataChannel('namedChannel', { reliable: false });
...
// register callback for handling remote data channels
pc.ondatachannel = handleChannel;
// handle events on the data channel
function handleChannel(dc) {
  dc.onerror = (error) => { /* handle error event */ }
  dc.onclose = () => { /* handle close event */ }
  // exchange application information with peer
  dc.onopen = (evt) => {
    dc.send(msg);
  }
  // act on data received by another peer
  dc.onmessage = (msg) => {
    console.log(msg.data);
  }
}
```

Listing 9: Message-oriented communication via a WebRTC data channel (Grigorik 2013, pg. 346)

# 5 Concept for a system supporting e-commerce fraud investigations

This chapter looks specifically into the concept of a collaborative system that will improve the situation described in the scenario in Section 3.5. To do this, the chapter discusses the overall concept of such a system on an high level, without going to much into implementation specific details. At the end, the chapter answers the question of what the system looks like, and what users should be able to achieve with it. In addition to these discussions, the chapter further looks into existing design approaches and analyses, why they are of no use for the specific scenario of this Master Thesis.

## 5.1 Collaboration on e-commerce fraud incidents

Based on the explanations in Chapter 3, and especially the scope definition for this Master Thesis in Section 3.5, the collaborative system for investigating e-commerce fraud incidents has to answer the central question:

*Is this transaction really a valid e-commerce transaction?*

Looking into the stakeholders, who can provide useful information to decide it, one will come up with:

1. **Merchants**, who can provide additional information of each e-commerce transaction in question.
2. **PSPs/issuers**, who have information about the credit card usage patterns and the original credit card owners.
3. **LSPs**, who can offer information about whether an order has already been shipped or not, and in the former case to whom it has been handed over.

Ideally, each of those participants would make parts of their internal databases available for the others to access and query for information in a shared information space. That would allow those stakeholders, who have to authorize or validate a suspicious credit card transaction, to analyse all available information as depicted in the Figure 5.1.

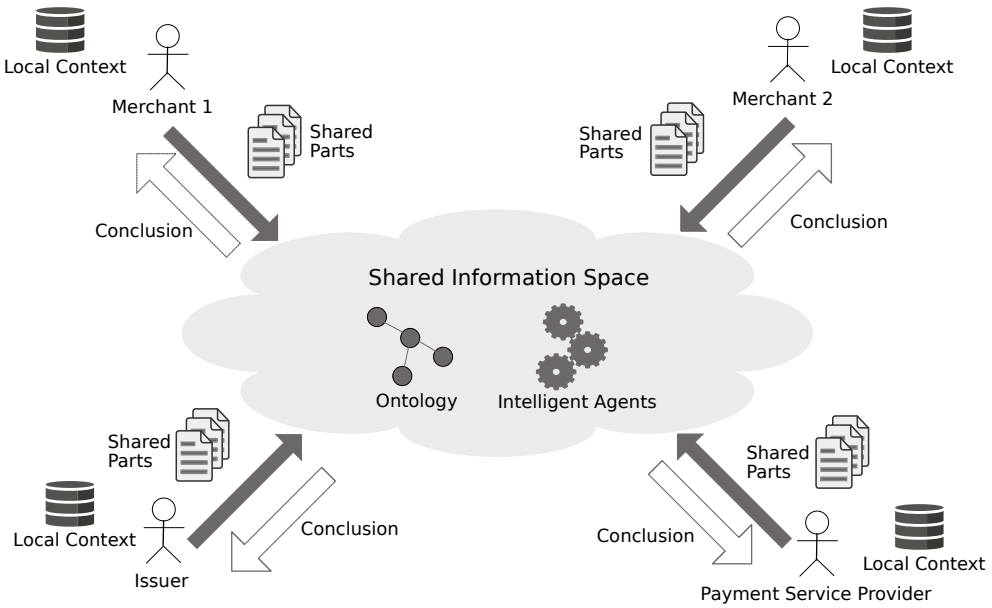


Figure 5.1: High-level concept of the system

In this figure one can see, how the relevant parties provide access to parts of their internal local context information within a shared information space. The collaborative system should allow participants to communicate and collaborate on the e-commerce fraud incidents from different places at the same time (see Section 4.1). Due to the fact that data from various sources have to be combined into a shared understanding of the e-commerce activities of a consumer, there is a need to harmonise and transform the information from each participant into a common data model to be able to analyse the combined data set. Based on the shared understanding of the e-commerce activities that have been done with a credit card recently, a set of intelligent agents (aka analysis tools) can assess them and present their findings, which can be valuable to any of the participants of the collaborative system.

## 5.2 An ER model for e-commerce transactions

Based on the analysis of the information each stakeholder holds and transmits to others in Section 3.2, the following ER model can be conducted for e-commerce transactions (see Figure 5.2). This figure shows not only the relevant information from the local contexts of each stakeholder, but also how they can be combined within a shared information space.

As the figure also shows there are *shared information tokens* that will be exchanged between various stakeholders. Those can be used in the collaborative system as a

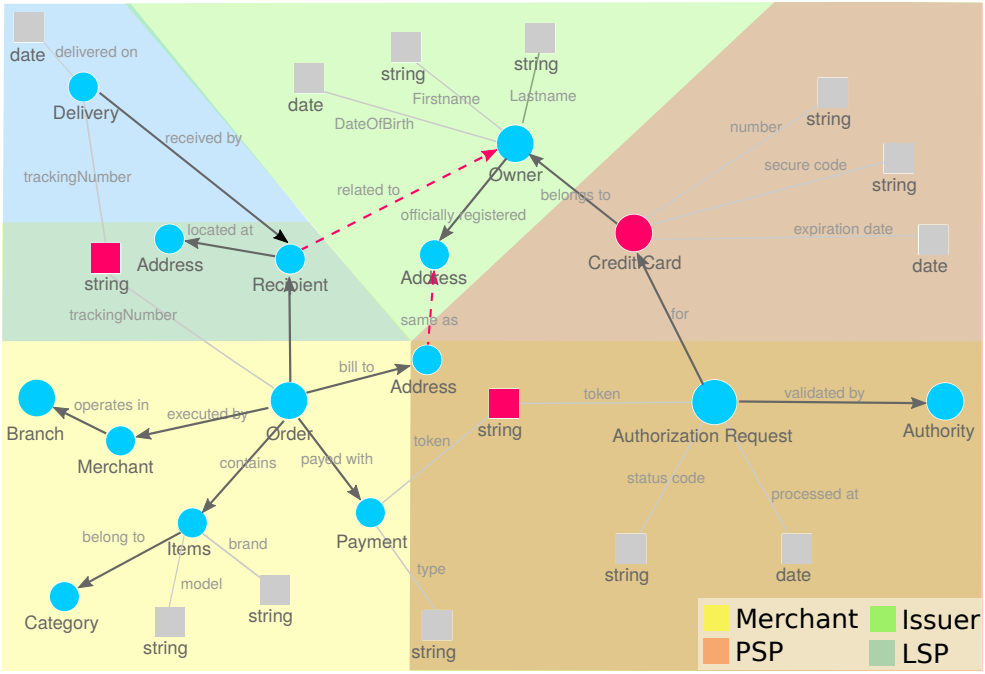


Figure 5.2: Entities and relations in the e-commerce scenario

reference for joining the distributed pieces of information into a combined view of an e-commerce transaction. There are actually three important tokens:

1. **Payment token**: shared between merchants and PSPs,
2. **Tracking number**: shared between merchants and LSPs,
3. **Credit card**: shared between issuers and PSPs

In addition to these tokens, Figure 5.2 also shows the important validation criteria. These are two important connections that have an influence on the decision whether an e-commerce transaction is evaluated as fraudulent or not. The two main criteria are:

1. **Billing address to owner address**: the billing address of the order has to match the registered address of the credit card owner
2. **Recipient to owner**: the recipient of the delivery has to be related to the owner of the credit card

Whereas the first criteria can be examined during the payment authorization process of an e-commerce transaction based on the information transmitted between merchants and PSPs or issuers, the second one is more difficult to validate (or can not be verified

at all). The only check the LSPs are able to do, before they are handing over the packaged items to the recipients, is to verify that they are the ones mentioned in the shipping address information of the order. If a recipient is somehow related to the owner of the credit card used for paying an order, or just a deceiver misusing a credit card, can not be confirmed by the LSP.

Also merchants, PSPs and issuers have no possibility to check for this criteria. Whereas the merchants are able to validate whether a consumer has send items to a shipping address before, they can not restrict consumers to choose only validated recipient addresses for their orders. Doing so would have negative impacts on the business success of the online merchant. The PSPs and issuers can not analyse this situation either, because both participants will not receive any information about the delivery address of an order with the payment authorization request from a merchant.

But just sharing the fact whether the shipping and billing address of an order is different or not between the relevant stakeholders is not enough. Although this information is necessary, it is not sufficient to make a decision about suspicious transactions. Other necessary information are whether the consumer has send orders to this shipping address before, and information about the content of the current order. Nevertheless, as mentioned in Section 3.5 looking at the transactions of just one of the online merchants is not sufficient either to solve the e-commerce fraud scenario of this Master Thesis. More sophisticated analysing capabilities are required for the collaborative system to be helpful for the e-commerce fraud investigation.

## 5.3 Analysis of e-commerce transactions

Based on the explanations in the previous sections the proposal is to link the transaction information from various merchants, LSPs, PSPs and issuers together into a shared information space to be able to analyse, if there are any orders that look extraordinary, and are likely not being made by the owner of the credit card to a certain extend. Thus, the collaborative system has to use statistical evaluations and probabilities to find and rate suspicious activities. Starting with the credit card in question, an issuer can query for the order details of all the transactions that have been recently done with the credit card online. To be able to do that, an issuer will likely have to query the PSPs for the payment tokens first, before asking the affected merchants for order details to any of those payment tokens. At the end each online transaction can be mapped into an ER model like the one shown in Figure 5.2, which resembles a large

graph of entities and their relations, and has the specific credit card in the centre of it. An abbreviated sample graph of this procedure can be seen in Figure 5.3.

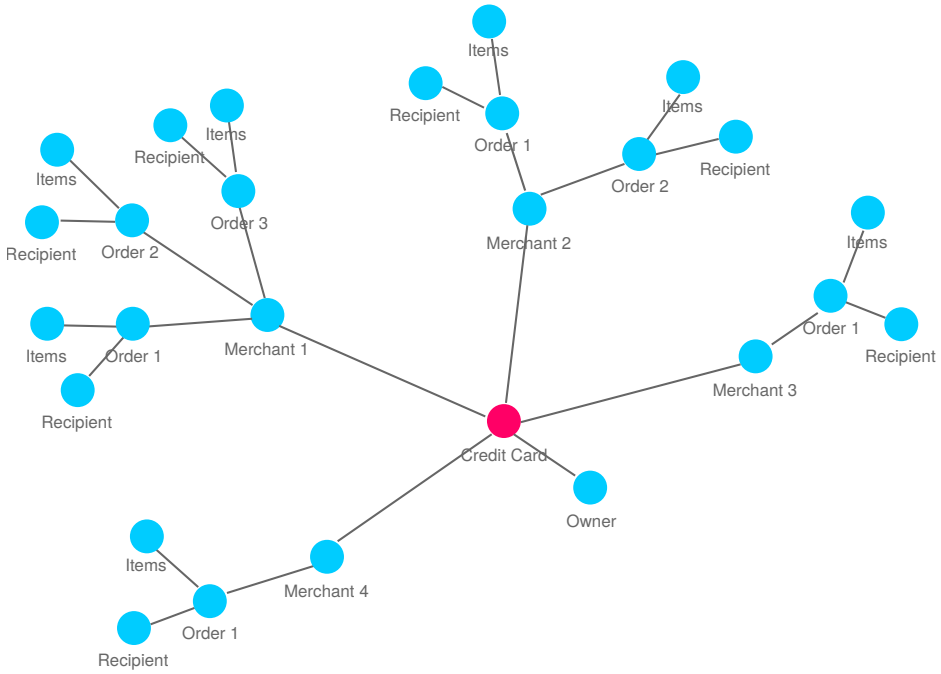


Figure 5.3: Building clusters of e-commerce transactions by merchant

As shown in this figure, the transactions will be organized by merchants first. But collecting the various order information into one combined data set is just the beginning of the e-commerce fraud incident analysis. Based on the information received, an issuer can already filter out transactions that have been shipped to different addresses than the one the credit card owner is registered for. Particularly for those edge cases it might be worth to ask for additional information from the affected merchants to be able to figure out, if the consumer has used one of these shipping addresses before. As a result the existing data set can be further enriched with supplementary transactional information from merchants at any time, if needed. In addition to the address information, an issuer can also analyse the item information (incl. category, brand and model) of each order to check for malicious activities.

But as already stated analysing the cluster of transactions on a merchant-by-merchant basis will not be sufficient to come up with a solid decision about a suspicious transaction. This is mostly due to the usage pattern of the fraudsters that have been explained in the scenario in Section 3.5. Based on this description, the various order details from

the merchants have to be mapped and linked against each other, so that the initial graph of transactions, which is organized by merchant, can be easily transformed into complementary representations, which use different criteria to cluster the transactions such as recipient addresses, branches of merchants, or product-related information.

This reshaping of the transactional details can lead to new insights about the "normal" shopping behaviour of a credit card owner, and can make deviations from this behaviour visible. By using a clustered graph for visualizing the combined data set on screen, the exploratory nature of knowledge generation and perception will be supported, and so this kind of representation can help speed up the investigation of e-commerce fraud incidents. An example visualization of a clustered graph, which groups information together based on a chosen criteria, is depicted in Figure 5.4. The different colours in this figure can represent different sources of information (e.g. e-commerce transactions from various merchants). In this example information that stands out from the "normal behaviour" can be found in the top right and lower left areas of the figure.

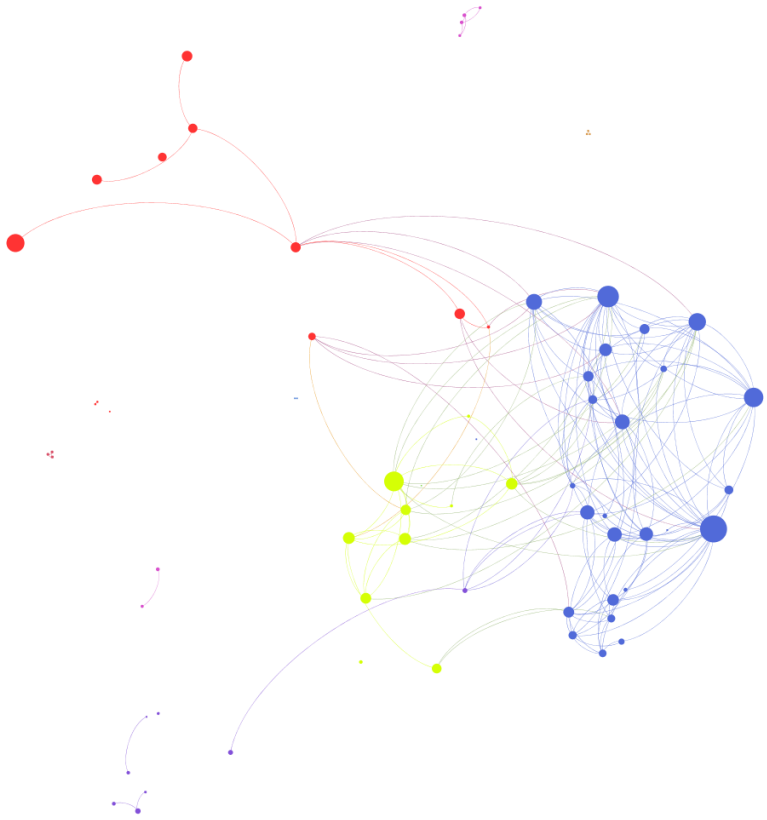

Figure 5.4: An example visualization of a clustered graph (Griffen 2012)

In addition to these clustered graph visualizations, the collaborative system can also support the e-commerce fraud investigation by switching the type of representation based on the chosen criteria; e.g. when clustering transaction details based on location information such as shipping addresses, the system can present the information as a heat map on a chart as is displayed in Figure 5.5.

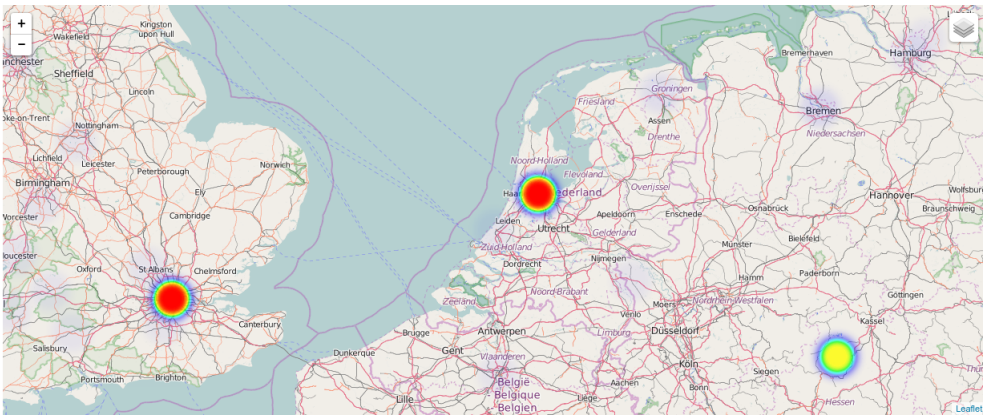


Figure 5.5: Heatmap displaying clusters of location-based information

Additionally, the collaborative system can provide a Hypertext-based visualization of the linked information to allow investigators navigating through the consolidated order details done with a credit card recently.

To sum up, the system have to support the collection and combination of e-commerce transaction information from various sources into a linked data set using a graph-oriented data model. This graph can further be analysed from multiple view points to validate, if there are any transactions that stand out from the "normal" shopping behaviour of the credit card owner. The starting point for the investigation is a sequence of payment tokens of recent credit card activities that an issuer can provide to the PSPs. The linked data set will initially collect and cluster the information from each merchant based on this list (see Figure 5.6). In case there are already suspicious information in one of these clusters, an issuer can ask for further details and enrich that specific cluster with additional order information for this consumer and that merchant. In the final step the system has to do the mapping and linking of the order detail information between each merchant to allow subsequent analysing and clustering of the transaction details based on various criteria.

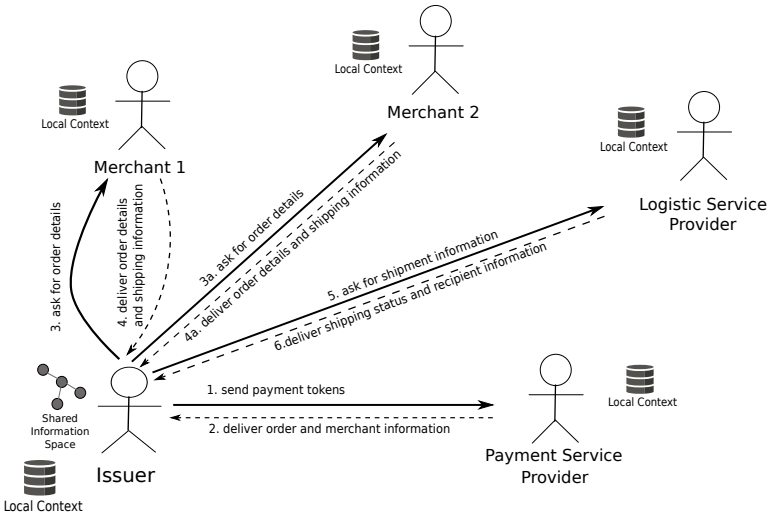


Figure 5.6: Information flow in the proposed collaborative system

## 5.4 Evaluation of existing design approaches

When trying to solve issues of information integration between organizations, there are already existing solutions that have to be examined whether they might fit the e-commerce fraud investigation scenario or not. This section looks into common existing approaches to collect and integrate information between IT systems.

### 5.4.1 ETL processes

To begin with, retrieving, transforming and combining data from multiple dispersed data sources is not a completely new problem, and is actually part of "Extract-Transform-Load" (ETL) processes *within* an organization. The basic idea is very much the same as in the proposal shown previously; namely to get as much information as possible from the various databases that are in use within a company, harmonise (aka transform) the data from each of them into a shared data model, and use the cleaned up and combined information repository for doing advanced business analysis and predictions later. Data within an organization is created and maintained by different business-related software tools. Each of these will usually store the information into their own database using a vendor-specific database schema. Other business-relevant information might be stored in structured files, sometimes using a proprietary format such as Microsoft Excel. Each of these data sources have to be accessed, the valuable information have to be extracted and mapped against each other, before the analysis of it can begin in a separate data store that holds the combined data set. The whole process is visualized in Figure 5.7.

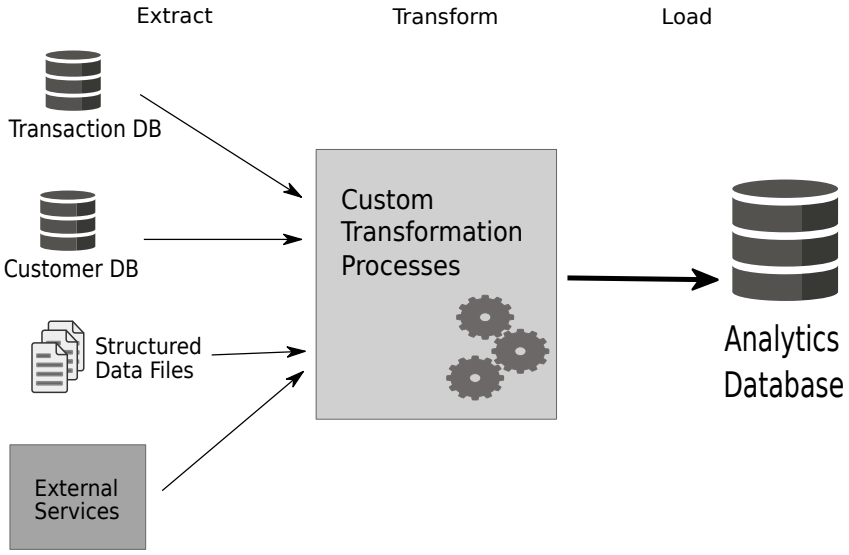


Figure 5.7: ETL processes within a company (Wood et al. 2014, pg. 165)

Although this description basically resembles the required activities as explained in the conceptual overview of the e-commerce fraud investigation system before, these ETL processes generally rely on an in-depth knowledge of the data structures that are used in each of the information sources, as well as require a direct access to the databases and files for retrieving the information. Although these preconditions are not cumbersome to work with *within* an organization, they are not suitable for situations, in which one has to integrate data sources across company boundaries. As the integration of the information takes place on the database level, granting external partners access to your internal databases will not only open up access to your business internals, but will also make it much more complicated to change the underlying database structures and business-related software tools later. Any changes to one of these would require elaborate negotiations between the owner of a data source and all of the external partners depending on it.

Beside these drawbacks, which make the ETL approach unsuitable for the e-commerce fraud investigation scenario as a whole, one can assume that these ETL processes are still in use for operating the daily business of each stakeholder involved. They can be helpful in the discussion later (see Section 6.2), when a decision has to be made about how each stakeholder can prepare and transform his internal data sources for external consumptions.

### 5.4.2 Web Services

With the development of the e-commerce scenario there was also a need to integrate business functionalities from various service providers on the Internet. Valid examples for these kind of integrations are the usage of the PSPs for doing the payment, as well as the LSPs for handling the shipping process. These approaches resulted in the "Service Oriented Architecture" paradigms, which enable application services provided by different vendors to talk to each other via a public facing programming interface (aka API). The only requirement for such interoperability to work properly is that each public interface follows some standardized or commonly agreed upon guidelines to be vendor-, platform- as well as programming language-agnostic. One possible implementation of these concepts are the so-called *Web Services*, which use the WS* protocols and standards from the W3C with the extensible markup language (aka XML) and the HTTP protocol at their core (Josuttis 2007).

Like the HTML format, which is used to represent Web pages on the Internet, XML is originally based on SGML, but instead of formalizing markup tags for structuring and styling textual content it is a meta-language allowing everyone to define their own markup languages. In this matter it doesn't dictate what tags are available to structure the information; instead it includes some basic guidelines for creating well-formed and valid documents that uses domain-specific tags, which can be freely defined and structured by the creator of the XML document. Therefore, it is better suited in situations, in which a computer has to parse and evaluate the content of a message; assuming the computer program knows the structure of that message. In an additional step, the author of the API could also specify an XML schema for each message, which describes the structure of the message with all the possible elements, their ordering, nesting level, and data types in detail. By doing so, the XML parser program can later verify the content of a message received against the XML schema, and validate if it is a valid document related to that schema definition. XML schemata are also expressed in XML format and have been standardized by the W3C.

Being able to create custom markup languages via XML has a huge benefit for machine-to-machine communication and is the basis for integrating Web Services (via the WS* protocols), but it still has limitations, when it comes to figure out the semantics of those XML messages. This is mostly due to the fact that each XML document represents a new markup language and needs a specific XML parser to be understood by the machine; additionally, to distinguish commonly used tag names in an XML document the creator has to place them into specific namespaces (aka XML namespaces). But

these XML namespaces further complicate the automatic processing of XML documents and increase the necessity to have custom instances of XML parsers for each XML document (Taylor & Harrison 2008).

An integration of information exchanged via Web Services is therefore handled separately for each Web Service interface. Looking at the payment service integration as *one* possible example, the following steps are necessary to allow a merchant to interact with the Web Service of a PSP:

- the PSP has to define and implement an interface (aka API) that a merchant can use for exchanging information,
- the API includes a set of request/response messages that hold the data being exchanged, usually specified in XML format, as well as a list of operations that the interface supports,
- the PSP has to document each of these messages and operations, incl. their intended structures and semantics,
- the PSP has to provide access to the API via an HTTP endpoint running on a server at a specific URL,
- the PSP usually restricts access to this interface for registered partners only; for doing so they have to provide a registration and identification mechanism,
- the merchant has to register with the PSP to be able to call into the Web Service API,
- the merchant receives some kind of token that can be used to identify with the Web Service later,
- the merchant has to implement an API-specific client-side wrapper that knows how to talk to the interface; incl. calling one of the available operations as well as serializing and de-serializing the messages, which will be transmitted between the Web Service and the client program,
- the client program from the merchant has to understand the structures and semantics of the messages exchanged with the Web Service and react on them accordingly.

Although other merchants, who want to use the same API from the PSP, can use the same client-side wrapper, which is sometimes also provided by that PSP for convenience, to be able to send/receive messages to/from this specific Web Service, they

still have to make the API-specific integrations into their Web shops. However, these integrations are only done in an *one-way* direction. To allow the merchants to provide information from their databases, the merchants have to do likewise and provide an API that others can use to query for information by following the same steps as mentioned above.

Additionally, as the structures and semantics of the messages and operations of each Web Service interface are not standardized, integrating with APIs from other PSPs or issuers result in doing the same integration steps again and again. To make things worse, the mapping and linking of the information coming from different APIs have to be implemented by each client to be able to analyse the combined data sets. It becomes clear soon that these necessary tasks will increase the time and efforts with each additional stakeholder, who wants to participate in the collaborative system, see Figure 5.8.

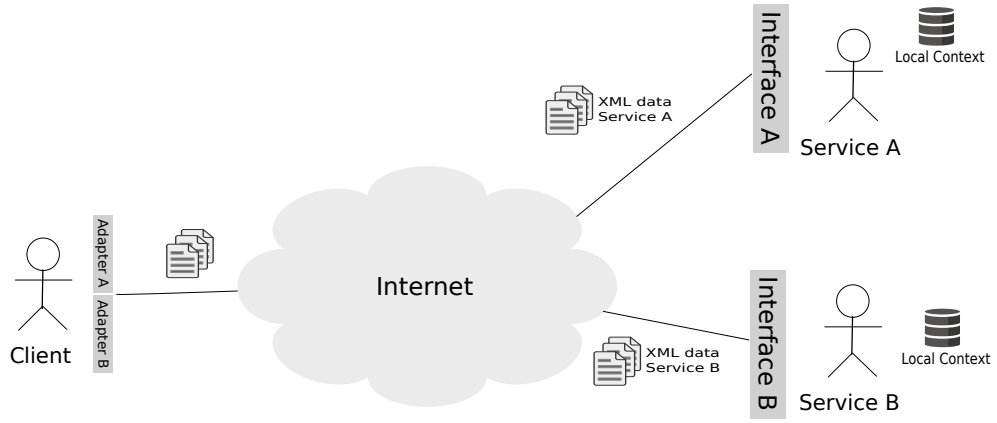


Figure 5.8: Data integration within the Web Service approach

As conclusion one could say that integrating information between a larger group of participants is very limited with the existing Web Service approach. The steps necessary for exchanging information result in huge efforts on all participating parties. As there is no common way to access and combine the information from each of the participants, beside using the fundamental HTTP protocol and XML data format, there have to be a lot of collaborative work between each of them upfront to come up with an approach for integrating the available APIs, and provide the rules for combining the different data structures. Due to these restrictions one can assume that an integration based on the Web Service approach will only work well with a limited number of participants. This might lead to a collaborative system that will only include larger online merchants, PSPs and issuers as participants, and therefore left out smaller companies from the e-commerce fraud investigation process. For a general solution of the

problem described in Section 3.5 this is not sufficient. Due to these limitations one will need other technologies that provide a better scalability and integration ability for the exchange of information between various, otherwise not strongly related organizations.

### 5.4.3 Semantic Web

"The Web is full of intelligent applications, with new innovations coming every day" (Allemang & Hendler 2011). But each of these intelligent Web applications are *solely* driven by the data available to them. Information that are likely coming from different places in the global information space, are usually *only* accessible via a custom API on the server hosting those resources (see Section 5.4.2). But the more consistent the information available to the smart Web application is, the better the service and its result will be. To support an integration of the data from various Web services, the semantics of the information delivered by each of them have to be available, and there has to be a generalized, formalized way to express the semantic of that data. The focus on a standard, which enables Web services to express the semantics of the data, also allows for global scalability, openness and decentralization that are the key principles of the World-Wide Web. The *Semantic Web* tries to give a solution for this problem by providing the Resource Description Framework (aka RDF) and related technologies (e.g. RDF schema, SPARQL, ...) for describing, linking and querying the data that a Web service delivers. But it doesn't reinvent the wheel; instead the Semantic Web builds upon existing, proven technologies such as XML, XML namespaces, XML schemata, and the URI to uniquely address resources on the Web (Allemang & Hendler 2011).

The main benefits of the Semantic Web approach are the specification of a standardized and generalized format to exchange information on the Web (aka RDF), as well as a commonly agreed way to access and query for them (aka SPARQL). The RDF data format does not only specify the syntax of the information exchanged, but also include the semantics (aka meanings) of them. Due to this fact, resources described in RDF format are consistent and semantically self-contained. These characteristics are achieved by providing information as a triple; that is a statement consisting of the resource in question (aka subject), a predicate and the specific value (aka object) for it. To be able to unambiguously identify the meaning of these statements, each part of such a triple is usually expressed with a unique URI. These URIs can be abbreviated via "prefix" definitions to make the whole statement easier to read (see also Section 4.2). To specify that there is an order "12345" from a "merchant1",

one can come up with the following RDF statement, which uses the Schema.org RDF vocabulary (Schema.org b) to describe an order:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix schema: <http://schema.org/> .
@base <http://www.merchant1.com/orders/> .

<12345>   rdf:type schema:Order;
          schema:orderNumber "12345"^^xsd:string .
```

Listing 10: An order specification in RDF

An RDF data set can contain one or more of such triples describing the resources of interest in detail. Usually these triples are visualized as directed graph, in that subjects and objects are displayed as nodes, and their predicates as edges between them. The order resource shown in Listing 10 above can also be visualized as directed graph as depicted in Figure 5.9.

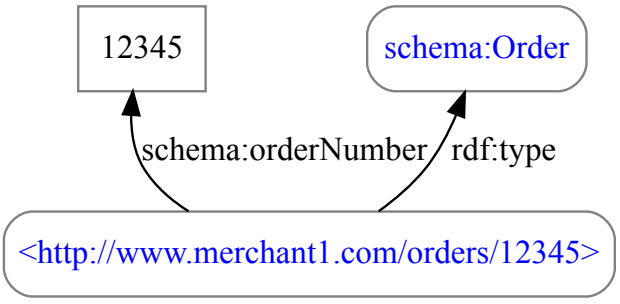


Figure 5.9: Graph-based visualization of the order from Listing 10

Additionally, the RDF format has build-in support for merging information from different data sources. This functionality is only working as expected, if the triples in the dispersed data stores are using the same URIs to refer to the same subjects or objects. In that situation, merging the triples from different RDF data sets will result in a locally linked data set holding the combined information as shown in Figure 5.10.

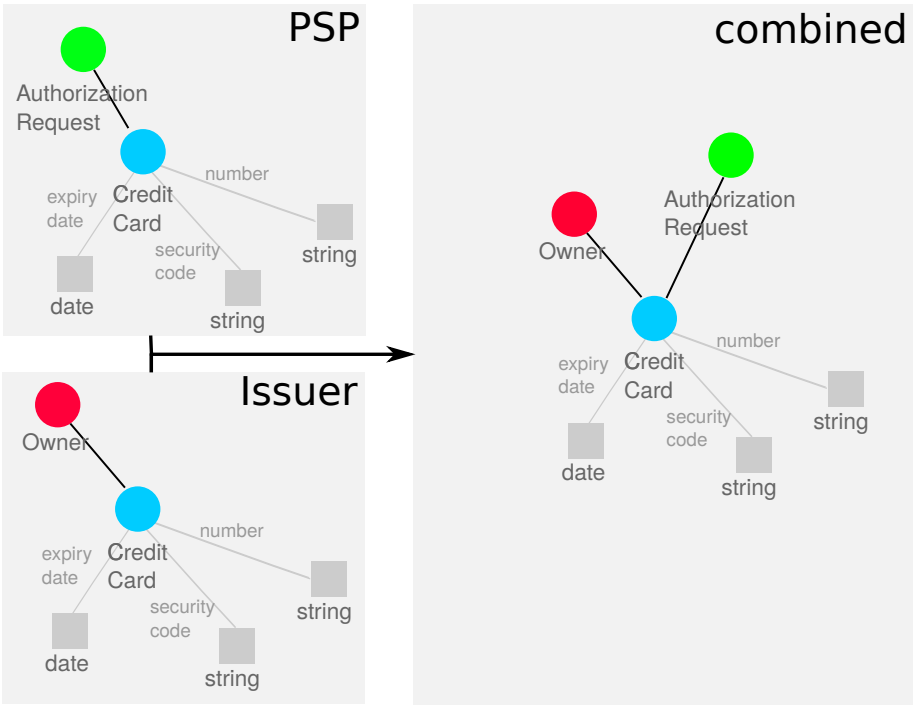


Figure 5.10: Combining two RDF files containing the same credit card entity

Beside being able to provide internal resources in an understandable RDF format for external consumption, the Semantic Web also specifies how to query and access these "information databases" on the Web. For that purpose the SPARQL protocol and query language has been defined. It does not only describes a language to query for information located in RDF data stores, but also specifies how to setup an HTTP endpoint on a server to make the RDF data set publicly available on the Internet.

Following the specifications of the Semantic Web standards, each relevant participant of an e-commerce fraud investigation system will have to transform the information from their internal databases into a set of triples with commonly agreed upon URI references and persist them into a RDF data store. For this transformation process an extension of the existing ETL processes in an organization can be used. Additionally, these RDF data sets will be made available publicly on the Web for information retrieval via the SPARQL protocol and query language. Each participant of the collaborative system will only need to know the specific addresses of these HTTP endpoints to be able to query them for information. The results of each query can be easily combined into a local RDF data set based on the merging capabilities of the RDF standard. This will decrease the efforts for integrating the data from various external

sources drastically. Also communicating with the different HTTP endpoints to access and query for information is being done in a much more efficient way based on the standardized SPARQL protocol and query language, see Figure 5.11.

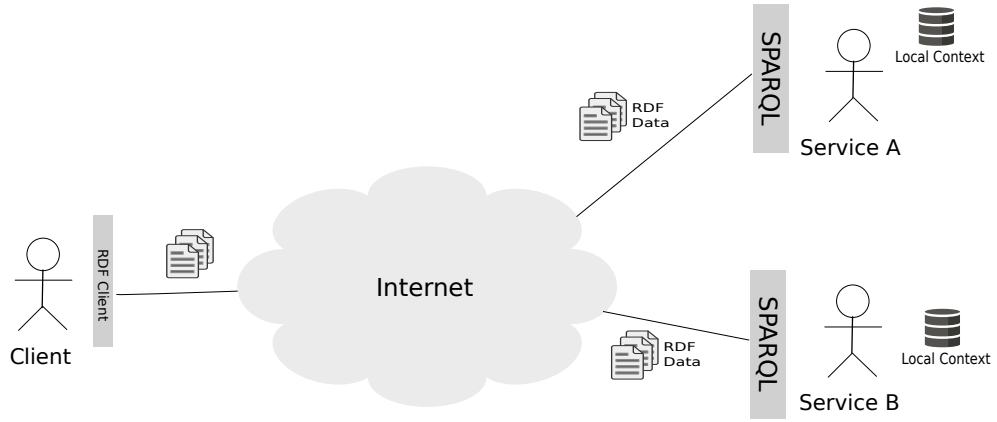


Figure 5.11: Data integration within the Semantic Web approach

As the underlying model of a RDF data set is resembling a graph-based data model, it will fit the concept of the proposed system from Section 5.3 perfectly. However, requiring every participant to setup and operate a publicly available SPARQL server will limit the use of this approach for the solution of the e-commerce fraud investigation scenario. As parts of the information that have to be exchanged between the relevant participants are confidential and/or business-critical, operating a public SPARQL endpoint on the Internet is a high security risk. Additionally, the SPARQL protocol and query language does not offer a way to restrict access to only a subset of the information in the RDF data stores. Any party, which is aware of the URL of a SPARQL endpoint, have access to all the information that are in the underlying RDF data stores, and can easily retrieve them with a single SPARQL query (see Listing 11). It is therefore no surprise that there are only a small set of publicly available SPARQL endpoints on the Internet — with the most commonly used one from DBpedia.org (DBpedia), which offers publicly available information from Wikipedia articles in RDF format.

As conclusion one can assert that the fundamental technologies of the Semantic Web standards are a good fit for exchanging and merging information between different stakeholders. But the usage of an all or nothing approach for querying the RDF data stores via the SPARQL protocol and query language is way to open for the e-commerce fraud investigation system.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?s ?p ?o    # select every subject-object-predicate triple found
WHERE {
 ?s ?p ?o          # do not specifying a condition returns everything
 }
```

Listing 11: Retrieving all information in an RDF store using SPARQL

## 5.5 Conclusion

To support the investigation of e-commerce frauds as described in Section 3.5, the collaborative system has to collect and combine transactional information from various online merchants of Web shops a credit card has been used with recently. The system has to support the combination and linking of the transactional details by utilizing a graph-based data model. Doing so will allow the system to classify and cluster the transaction information based on various criteria, which can help the investigator to figure out abnormal behavioural patterns in the credit card usage on the Internet. Visualizing the combined data set can make use of the graph-based data model and present the transaction details as a clustered graph on screen. Additionally, the representation of the information can be adapted to the requirements of the investigators.

As the previous section showed in detail, existing approaches are of limited use for the collection and combination of dispersed transactional details in this scenario. The leading approach for the e-commerce fraud investigation system will have to combine the best characteristics from the Web Service and the Semantic Web designs.

As for the Web Service approach the most valuable aspects of it are:

- access to the HTTP endpoints can be limited to a certain set of communication partners,
- these partners have to authenticate with each Web Service first,
- based on the identification of the partners only certain aspects of the information can be exchanged, and execution of Web Service operations can be restricted.

When looking at the Semantic Web approach it's most beneficial functionalities are:

- providing information in a semantically self-contained way,

- the ability to merge and link together information from different RDF data stores locally,
- the graph-based data model underlying the RDF data stores,
- the usage of SPARQL to query and analyse the combined data set locally.

In the following Chapter 6 the Master Thesis will come up with an approach that uses the fundamental technologies from the Semantic Web for information sharing and integration, as well as peer-to-peer communication technologies for securing and restricting access to the RDF data sets for relevant participants of the e-commerce fraud investigation scenario only.

# 6 Design of a collaborative system

This chapter is about the design of a collaborative system that supports the investigation of e-commerce fraud incidents. It starts with a discussion of the semantics of the underlying RDF data sets, and how these can be combined across various organizations. After that, it shows how these information can be provided by the relevant participants based on the e-commerce fraud investigation scenario described in Chapter 3. For this purpose, it looks in detail into the partially centralized P2P communication architecture, and shows how that can be used for securely sharing the relevant information between the stakeholders.

## 6.1 RDF vocabularies and Web Ontologies for e-commerce

As a major objective of the e-commerce fraud investigation system is to collect the various transactional information from online merchants, PSPs and issuers, combine and link them together, as well as analyse the resulting data set from different view points to find abnormal activities, the information exchanged between the relevant participants either have to follow commonly available RDF vocabularies, have to be based on a custom shared RDF vocabulary that has been specifically designed for this system, or have to be mapped and linked against each other from different RDF schema specifications.

### 6.1.1 Reuse of common RDF vocabularies

One valid approach to come up with a data schema for the collaborative system is to take a look into commonly used RDF vocabularies and Web ontologies, and try to figure out whether they can be used for describing the information that need to be exchanged between participants of the e-commerce fraud investigation system. When consulting the Semantic Web community for commonly agreed upon and highly used RDF schema specifications, one will come up with the following list (see Table 6.1):

| Name | Prefix | Describes | Namespace URI |
|---|---|---|---|
| Dublin Core | dc: | Meta data | http://purl.org/dc/terms/ |
| FOAF | foaf: | People | http://xmlns.com/foaf/0.1/ |
| Geo | pos: | Positions | http://www.w3.org/2003/01/geo/wgs84_pos# |
| Geo Names | gn: | Locations | http://www.geonames.org/ontology# |
| Good Relations | gr: | Products | http://purl.org/goodrelations/v1# |
| RDF | rdf: | Core framework | http://www.w3.org/1999/02/22-rdf-syntax-ns# |
| RDFS | rdfs: | RDF vocabularies | http://www.w3.org/2000/01/rdf-schema# |
| Schema.org | schema: | Schema.org vocabularies | http://schema.org/ |
| SKOS | skos: | Controlled vocabularies | http://www.w3.org/2004/02/skos/core# |
| vCard | vcard: | Business Cards | http://www.w3.org/2006/vcard/ns# |
| Web Ontology Language | owl: | Ontologies | http://www.w3.org/2002/07/owl# |
| XML Schema Datatypes | xsd: | Data types | http://www.w3.org/2001/XMLSchema# |

Table 6.1: Commonly used RDF vocabularies on the Web (Wood et al. 2014, pg. 41)

Based on these existing schema specifications describing a fictive consumer named "Max Mustermann" incl. his home address can be done by combining data utilizing the FOAF and vCard vocabularies into a RDF data set such as described in Listing 12 and visualized as directed graph in Figure 6.1. The described resource is uniquely identified by the URI `http://www.merchant1.com/customers/MaxMustermann`. Additionally, one can see that these vocabularies use expressive names for their entities and predicates, which make it easier to understand their intended meanings (e.g. "foaf:givenname", "vcard:locality", ...).

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix vcard: <http://www.w3.org/2006/vcard/ns#> .
@base <http://www.merchant1.com/customers/> .

<MaxMustermann> rdf:type foaf:Person;
                rdfs:label "Max Mustermann";
                foaf:family_name "Mustermann";
                foaf:givenname "Max";
                foaf:gender "Male";
                foaf:title "Mr.";
                vcard:adr [
                 rdf:type vcard:Home;
                 vcard:street-address "Mustermannstr. 12";
                 vcard:locality "Musterstadt";
                 vcard:region "North-Rhine Westfalia";
                 vcard:postal-code "33123";
                 vcard:country-name "Germany"
                ] .
```

Listing 12: Personal related information about a fictive consumer in RDF

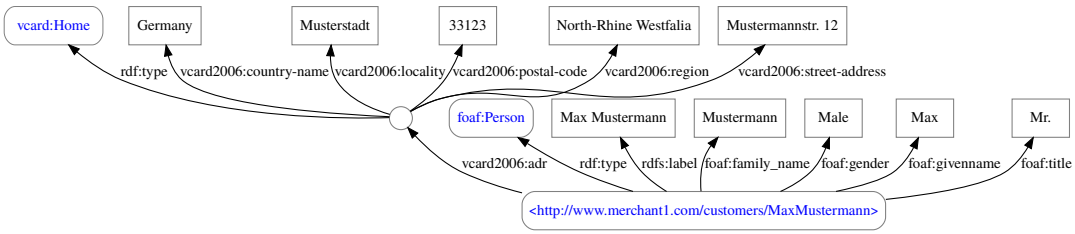


Figure 6.1: Graph representation of consumer information from Listing 12

However, being able to describe persons and their addresses is just a subset of the entities and relations found in the e-commerce scenario. When looking back to the initial ER model of an e-commerce transaction as shown in Section 5.2, one can map the entities that are currently available in the e-commerce scenario to the existing RDF vocabularies such as follows (see Table 6.2):

| **Information** | **RDF vocabulary** |
|---|---|
| Consumer | FOAF |
| Credit Card Owner | FOAF |
| Billing Address | vCard |
| Shipping Address | vCard |
| Location Information | Geo Names |
| Merchant | GoodRelations |
| Items | GoodRelations |
| Item Categories | GoodRelations |
| Brands | GoodRelations |
| Payment Types | GoodRelations |

Table 6.2: Possible usage of RDF vocabularies for e-commerce transaction information

As this table shows, there are some parts of the e-commerce ER model that can be expressed with existing RDF vocabularies extensively such as personal related information via FOAF and vCard, whereas other parts can not be stated in-depth (e.g. credit card information), or are not specified at all (e.g. tracking of the delivery). Due to these circumstances, one usually have to build an own RDF vocabulary or Web ontology that fills in the missing pieces, and refers to the existing concepts whenever appropriate.

When trying to model the information of a credit card as displayed in Figure 5.2, a possible result will be the RDFS specification shown in Listing 13. This definition of a credit card resource explicitly reuses entities from the FOAF and GoodRelations ontologies by defining that:

- the owner of a credit card has to be of type “Person” from the FOAF ontology,
- the type of a credit card has to be an instance of the type “PaymentMethodCreditCard” from the GoodRelations ontology.

As most of the parts of the e-commerce data model shown in Figure 5.2 can not be expressed with the existing RDF vocabularies directly, filling in the gaps would mean to come up with a large set of custom entities and relationships, which will limit the usage of the system as explained in Section 6.1.2.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix gr: <http://purl.org/goodrelations/v1#> .
@base <http://www.example.com/ecommerce#> .
# define the subject "CreditCard"
<CreditCard>      rdf:type rdfs:Class;
                  rdfs:comment "Describes a credit card in the
  ↪  E-commerce scenario";
                  rdfs:label "A credit card" .
# define the property "ExpirationDate" on subject "CreditCard"
<ExpirationDate>  rdf:type rdfs:Property;
                  rdfs:domain <CreditCard>;
                  rdfs:range xsd:date;
                  rdfs:label "Expiration Date" .
# define the property "SecureCode" on subject "CreditCard"
<SecureCode>      rdf:type rdfs:Property;
                  rdfs:domain <CreditCard>;
                  rdfs:range xsd:string;
                  rdfs:label "Security Code" .
# define the property "Number" on subject "CreditCard"
<Number>          rdf:type rdfs:Property;
                  rdfs:domain <CreditCard>;
                  rdfs:range  xsd:string;
                  rdfs:label "Credit Card Number" .
# define the property "BelongsTo" on subject "CreditCard"
<BelongsTo>       rdf:type rdfs:Property;
                  rdfs:domain <CreditCard>;
                  rdfs:range <foaf:Person>;
                  rdfs:label "Credit Card Owner" .
# define the property "Type" on subject "CreditCard"
<Type>            rdf:type rdfs:Property;
                  rdfs:domain <CreditCard>;
                  rdfs:range <gr:PaymentMethodCreditCard>;
                  rdfs:label "Type of Credit Card" .
```

Listing 13: A specification for a credit card in RDFS

When analysing the list of existing and actively used RDF vocabularies and Web ontologies in Table 6.1, one will also find the Schema.org initiative (Schema.org b). This meta data vocabulary was initially designed by the leading search engines (e.g. Google, Microsoft and Yahoo!) to allow authors of Web sites to markup their HTML documents in a way, so that they are better understood by these search engines. The Schema.org vocabulary is actively maintained by its community, includes new concepts with each release, and also offers an extension mechanism to implement additional vocabularies with terms that are not part of the core specifications (Schema.org a) yet. In one of the past releases the maintainers also introduced all of the existing concepts of the GoodRelations ontology into the Schema.org vocabulary (Guha).

As online merchants likely provide semantic meta data for their products and offerings in the vocabulary of Schema.org with the objective to improve their listings on search engine results (also known as SEO) already, one can reuse parts of these information for the e-commerce fraud investigation scenario. Additionally, the wide-ranging scope of aspects declared in the Schema.org vocabulary can make it a good fit for the collaborative system to support the investigation of e-commerce frauds. When trying to map the initial ER model from Section 5.2 to the Schema.org core specifications, one will basically come up with a schema as displayed in Figure 6.2.

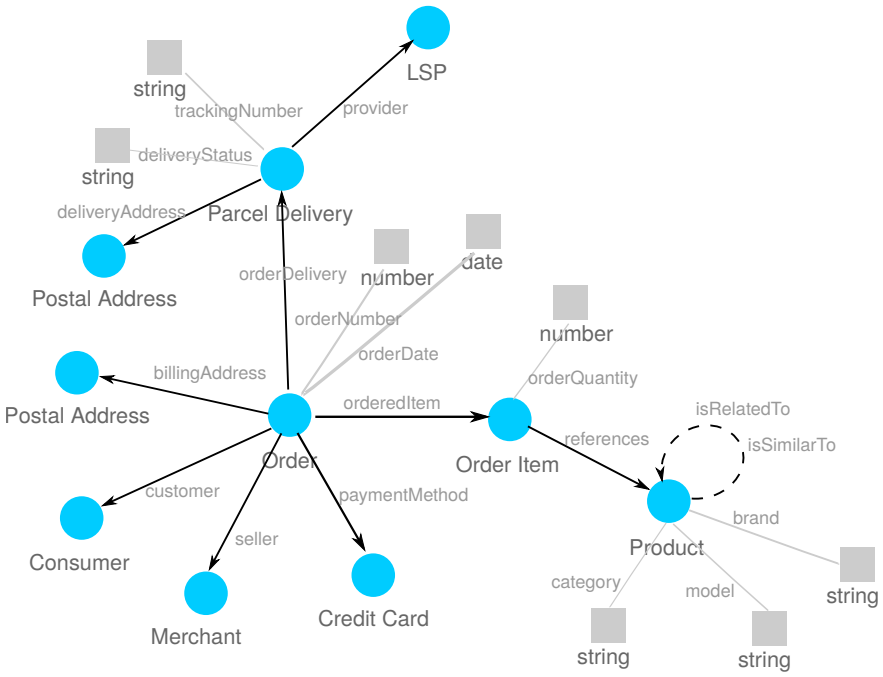


Figure 6.2: Schema.org based mapping of an e-commerce transaction

### 6.1.2 Creation of a custom RDF vocabulary

Another possible approach to harmonise information in the collaborative system is to define a completely new RDF vocabulary or Web ontology for the proposed e-commerce fraud investigation system, and share that with every possible stakeholder. This specification will have to define all the entities and relations known to the collaborative system and describe them in RDFS format (see Section 4.2).

A major drawback of this approach is that new participants of the system will have to implement the conversion of their internal data structures to a RDF data set that follows the predefined schema definition first, before even being able to join in. This will limit the general usage of the collaborative system, and will therefore not further considered in detail.

### 6.1.3 Mapping of RDF vocabularies

Although it is possible to model an e-commerce transaction solely with the Schema.org specifications as shown in Figure 6.2, the collaborative system likely has to take care of the mapping of the transactional information coming from various sources to be able to combine them later. As the Semantic Web does not restrict how organizations structure and express their information, and due to the "AAA slogan" (see Section 4.2), there are likely different RDF representations of an e-commerce transaction in-use and have to be brought together.

The W3C standards for the Semantic Web also include support for these mapping issues, because they will also come up when trying to combine semantic information available around the Web. The following axioms are available in the RDFS and OWL specifications explicitly for that purpose:

- **rdfs:subClassOf:** a relation of type "rdfs:subClassOf" defines a specialization of a class, in which the child class inherits all the properties of the parent class,
- **rdfs:subProperyOf:** a relation of type "rdfs:subPropertyOf" defines a specialization of a property, in which the child property inherits all constraints of the parent property,
- **owl:equivalentClass:** a relation of type "owl:equivalentClass" specifies the equality of classes coming from different RDF vocabularies or Web ontologies,
- **owl:equivalentProperty:** a relation of type "owl:equivalentProperty" specifies the equality of classes coming from different RDF vocabularies or Web ontologies

If a merchant wants to state that a product related information, which is delivered as resource using the GoodRelations vocabulary, is equal to product information that can be found in the Schema.org specification, he or she can do so as follows (see Listing 14):

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix schema: <http://schema.org/> .
@prefix gr: <http://purl.org/goodrelations/v1#> .

# mapping classes and properties between GoodRelations and Schema.org
gr:ProductOrService owl:equivalentClass schema:Product .
gr:category owl:equivalentProperty schema:category .
gr:color owl:equivalentProperty schema:color .
gr:description owl:equivalentProperty schema:description .
gr:hasBrand owl:equivalentProperty schema:brand .
gr:hasEAN_UCC-13 owl:equivalentProperty schema:gtin13 .
gr:hasGTIN-14 owl:equivalentProperty schema:gtin14 .
gr:hasGTIN-8 owl:equivalentProperty schema:gtin8 .
gr:name owl:equivalentProperty schema:name .
[...]
```

Listing 14: Mapping product-related information from GoodRelations to Schema.org

These mapping statements from one RDF vocabulary to another can be either created and injected into a RDF data store by the party, who is going to merge information from different sources according to needs, or can also be part of the resource specification coming from an external source. In the former case the stakeholder, who is collecting and combining the information from various sources, has to maintain the additional triples to map information between each RDF vocabulary used, and include them in the processing of the external statements within the combined RDF data store. With an increased number of participants, which are using disjunct RDF vocabularies, the effort and time to manage and create these mapping instructions on the collectors side will increase tremendously. Thus, the second approach is the preferred one. In that situation the RDF description of an entity coming from an external source is already stating the mapping to one or more well-known RDF vocabularies (e.g. the Schema.org specification mentioned above). This will reduce the effort and time to combine the

information from different sources, and will only slightly increase the effort on the side of the external partner to prepare their internal information for external consumptions.

## 6.2 Use of RDF data sets for e-commerce fraud investigation

With these methods in place, one can now specify how the relevant participants have to provide their information, so that they can be combined and analysed in the collaborative system. This section explains how the different participants might prepare their local context information for external consumption, and how the transactional details from various online merchants can be combined on the level of individual resources to be able to analyse and cluster the transactions by different criteria as described in Section 5.3.

### 6.2.1 Preparation of information

As explained in Section 5.4.1, there are likely ETL processes in-use within the IT operations of every stakeholder. These processes are usually collecting and combining information from internal data sources for *internal* business analysis, but can also be used to prepare internal data for external consumption. In the latter case, the parts of the relevant information for the e-commerce fraud investigation have to be extracted from the internal databases and encoded in a RDF data set incl. the RDFS vocabulary used and any required mapping statement to well-known vocabularies such as the one from Schema.org as shown in Figure 6.2.

#### Merchant

The merchants should be able to provide RDF encoded information for their orders based on a given payment token, or based on a consumer identification. The former selector is likely used in the initial phase that collects all required information of orders, which have been done with a credit card recently. The latter one is of interest, if there are suspicious transactions found for a consumer, and the merchant will have to provide additional order details for that individual.

The merchants provide the following information in RDF:

- **Product related information:** as part of the order details the merchants have detailed information about the products that have been bought by a consumer. These information include the brand, model, as well as product categories of each item within an order. These are the attributes that are likely of interest

in the e-commerce fraud investigation. If merchants have these information in a RDFa format on their Web sites already, they can refer to those data via the "rdfs:seeAlso" predicate, which holds a URI to an external resource that contains additional information for the subject (see Listing 15 for an example). Additionally, the product related information might be available in different languages on the Web shop. The merchants should use the English expression for each textual identifier in a RDF data set, and express language-dependent terms via the "rdfs:label" predicate that is used for a human-friendly name of the resource and supports language specifiers (see Listing 16 for an example),

- **Consumer related information:** as part of the order details the merchants also have the personal related information of the buyers incl. their billing and shipping addresses,
- **Merchant related information:** the merchants can also provide information about themselves, such as the retail branch they are operating in.

```
1  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
2  @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
3  @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
4  @prefix schema: <http://schema.org/> .
5  @base <http://www.merchant1.com/orders/> .
6
7  <O12345>  rdf:type schema:Order;
8            schema:orderedItem [
9              rdf:type schema:Product;
10             schema:name "Self-cleaning refrigerator";
11             rdfs:seeAlso <http://www.merchant1.com/catalog/P12345> .
12           ] .
```

Listing 15: Specifying a link to a Web site for looking up product-related information in RDF[1]

[1]Please note that the application has to resolve the URI and embed the external RDF data set at this position.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix schema: <http://schema.org/> .
@base <http://www.merchant1.com/catalog/> .

<P12345>  rdf:type schema:Product;
          schema:name "Self-cleaning refrigerator";
          rdfs:label "Selbstreinigender Kühlschrank"@de;
          rdfs:label "Self-cleaning refrigerator"@en;
          rdfs:label "refrigerador autolimpiable"@es .
```

Listing 16: Specifying a product with labels in three different languages in RDF[2]

#### Payment Service Provider

The PSPs provide information about a payment token and the authorization request that belongs to it. These information are required to link the credit card to the order information from a merchant. Thus, the PSPs can act as a broker between the issuers and online merchants. On the one hand they have a strong business relationship with the issuers for any payment related activities, and on the other hand they have an integration of their Web Service APIs at the merchants (see Section 5.4.2). The benefit of this is that issuers do not have to know about any online merchant available on the Internet, because they can get contact information to any of these merchants from the PSPs. In a distributed P2P communication scenario the PSPs can also use the RDFS predicate “rdfs:seeAlso” for providing a link to the online merchant affiliated to a payment authorization in their RDF data set as shown in Listing 17.

[2]Please note that the name of the product has been stated without a language specifier, which makes it valid globally.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix schema: <http://schema.org/> .
@base <http://www.payservices.com/payments/> .

<C12345>  rdf:type schema:PayAction;
          schema:recipient [
            schema:name "ACME Corp.";
            rdfs:seeAlso "http://www.acme.com/about"
          ] .
```

Listing 17: Linking to an online merchant in the RDF from a PSP

#### Logistic Service Provider

The LSPs provide information about a tracking number and the delivery status of an order. If the recipients have to show their ID card, or have to place a signature on the delivery receipt, the LSPs can also hand over personal related information about them. The information will be requested based on the tracking number that is shared with a merchant. Thus, the merchants will be acting as a broker between the issuers and the LSPs due to the existing business integrations between merchants and LSPs.

#### Issuer

The issuers holds information about credit cards and their owners incl. personal related information. They are the ones, who are usually initiating the e-commerce fraud investigation by asking the PSPs for detailed order information to a payment authorization request. They will also collect all the information from the different stakeholders, and have to combine and analyse them to be able to assess a credit card transaction. To be able to do so, they will have to use a RDF data store, in which the dispersed RDF data sets are imported and linked against each other (see next section).

### 6.2.2 Linking of information from various sources

The issuers must be able to link together the transactional details from various online merchants on the level of individual resources for analysing and clustering the order details on different aspects. This can be achieved by either assigning unique IDs to important subjects, or by inferring equality based on the RDF schematas used.

#### Uniquely identify entities in a RDF data set

As explained in Section 5.4.3, the build-in merging capabilities of the RDF specification will rely on unique URIs, which are used to describe the same entities found in different RDF data sets. Thus, individual resources can be extended with the "owl:sameAs" predicate from the OWL specification to provide a URI that unambiguously define the subject (see Listing 18 for an example).

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix schema: <http://schema.org/> .
@base <http://www.merchant1.com/customers/> .

<MaxMustermann> rdf:type schema:Person;
                schema:name "Mustermann, Max";
                rdfs:label "Max Mustermann"@en;
                schema:adress [
                  rdf:type schema:PostalAddress;
                  schema:addressLocality "Cologne";
                  rdfs:label "Cologne"@en;
                  owl:sameAs <http://dbpedia.org/resource/Cologne>
                ] .
```

Listing 18: Specifying a link to a DBpedia resource to uniquely identify an entity in RDF

When looking at the e-commerce ER model as defined in Section 5.2, the following information must be uniquely identified within the RDF data sets coming from the participants of the collaborative system:

- **Personal related information** such as consumers, recipients and owners of credit cards,
- **Location based information** such as billing and shipping addresses, as well as the location a credit card owner is registered for,
- **Product related information** such as categories, subcategories, brand, model and product identifier,
- **Merchant related information** such as the branch of a merchant.

To support the unique identification of entities in the RDF data of an e-commerce transaction, one can either refer to publicly available RDF data sets on the Internet such as GeoNames (GeoNames) or DBpedia (DBpedia), or use unique identification scheme (aka URN) such as phone numbers, e-mail addresses, or EAN codes.

The publicly available RDF data sets can provide unique URLs for locations and named places, as well as brands and item categories. A product related RDF data set, which was available in form of the ProductDB initiative until recently (Bouzidi et al. 2014), has been shut down. Thus, the linking of products can no longer be done by referencing unique URLs from the Web, but will have to use the global trade item number (aka GTIN) of each product instead. Additional aspects of an item such as brand, categories and subcategories can eventually be found on DBpedia (DBpedia).

A problem that will come up is the unique addressing of personal related information such as uniquely identifying a consumer. The collaborative system can not rely on linking the personal related information based on properties such as “familyName” and “givenName” alone (see Listing 12). There could be typos in the information coming from various RDF data sets, and distinct individuals can still have the same naming information. One possible approach to bring personal related information together, is to link them based on the e-mail address of an individual. An e-mail address is a globally unique addressing scheme, and one can assume that two entities, which are using the same e-mail address, are referring to the same subject. However, this is only a weak link as an individual can have more than one e-mail address, and could use varying e-mail addresses for online shopping trips at different merchants. Additionally, fraudsters can fake e-mail addresses, or can break into the mail system of a consumer with the intent of doing an identity theft. Therefore, a more sophisticated linking algorithm for personal related information is needed in the collaborative system. This procedure may take into account a combination of “familyName”, “givenName”, “dateOfBirth”, as well as location based information such as the position of the computer used regularly by a consumer to uniquely identify an individual.

To sum up, the unique identification of important entities from an e-commerce transaction can be based on the following aspects (see Table 6.3):

| Entity | Unique Identifier | Public Data Set | Example |
|---|---|---|---|
| Person | evt. e-mail address, combination of various attributes | n/a | `mailto: max.mustermann@t-online.de` |
| PostalAddress | Location, Position | Geo Names | `http://sws.geonames.org/2886242/` |
| Item | GTIN, ISBN | n/a | `gtin:9781617290398` |
| Brand | Name | DBpedia | `http://dbpedia.org/resource/Samsung` |
| Organization | Web Site URL | n/a | `http://www.samsung.com` |

Table 6.3: List of possible criteria to uniquely identify entities of an e-commerce transaction

#### Infer equality in a RDF data set

In addition to the unique identification scheme for entities in dispersed RDF data sets, the OWL specification includes axioms that allow to infer equality into a combined RDF data set. The two important predicates for that are "owl:FunctionalProperty" and "owl:InverseFunctionalProperty" (see Section 4.2.3). A possible use case for that in the e-commerce scenario is the linking of product related information. As a product is uniquely identified by its global trade item number, a predicate for this attribute should be classified as "owl:InverseFunctionalProperty". By doing so, the reasoners working on a combined RDF data set can infer that a subject A is equal to another subject B, if both subjects have the same predicate-value statement as shown in Listing 19.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix schema: <http://schema.org/> .
@prefix gr: <http://purl.org/goodrelations/v1#> .

# specify that predicate gtin13 & hasEAN_UCC-13 are
  ↪ InverseFunctionalProperties
schema:gtin13 rdf:type owl:InverseFunctionalProperty .
gr:hasEAN_UCC-13 rdf:type owl:InverseFunctionalProperty .

# specify that both predicates are equal
schema:gtin13 owl:equivalentProperty gr:hasEAN_UCC-13 .

# describe product in Schema.org vocabulary
<P12345>  rdf:type schema:Product;
          schema:name "Self-cleaning refrigerator";
          schema:gtin13 "111-222-333-444-5" .

# describe product in GoodRelations vocabulary
<P556677> rdf:type gr:ProductOrService;
          gr:name "Self-cleaning refrigerator";
          gr:hasEAN_UCC-13 "111-222-333-444-5" .

# a reasoner will infer this additional statement for the products
<P12345> owl:sameAs <P556677> .
```

Listing 19: Infer equality between two product specifications using different RDF vocabularies

After being able to infer that both products are the same when their global trade item number are equal, additional statements can be deduced — e.g. as a product identified by a GTIN generally relates to just one brand, as well as to one model, the RDF data set can further specify that these predicates are of type “owl:FunctionalProperty”. This would allow the reasoner to inject additional “owl:sameAs” statements into the combined RDF data set.

## 6.3 A partially centralized P2P system proposal

In the e-commerce fraud scenario, which has been selected for this Master Thesis in Section 3.5, the issuer of a credit card is the participant, who initiates a collaborative session to investigate a suspicious online transaction. Issuers are recognizing the active use (and maybe misuse) of a credit card in the online and the offline world first, and are also getting notifications about any suspicious activity from their fraud prevention systems. Due to these circumstances, a valid design proposal for the e-commerce fraud investigation system is based on a partially centralized P2P communication network, in which the issuer of a credit card is taking over a leading position.

### 6.3.1 Role of the issuer

In case a suspicious activity has been detected by the fraud prevention system of an issuer, one of their investigators will initiate a collaborative session with the relevant stakeholders that have been selected based on the usage history of the credit card in question. To establish the P2P communication session, the investigator creates a new WebRTC communication session in the collaborative system, which has been implemented as a Web application. By doing so, the investigator receives a unique session ID that can be transmitted to merchants, PSPs and LSPs, so that they can join the session in the collaborative system. As the investigator is just being aware of the PSPs affected, they can invite them directly due to the existing business relationship between both parties. Based on the given credit card information, the PSPs can relate each payment authorization request to an order of an online merchant by referring to the payment authorization token. So the PSPs will know the merchants that have to be involved in this e-commerce fraud investigation, and can hand over contact information to the issuer for inviting them. Each online merchant can do likewise with the LSP that has been used to handle the delivery of an order via the tracking number (as depicted in Figure 5.6).

As soon as the relevant merchants, PSPs and LSPs have joined the P2P communication session, they will start sharing their information with the issuer. The information exchanged have been prepared as RDF data sets by internal ETL processes, and has been made available to the support staff of each stakeholder. During the information sharing process the RDF data sets of each stakeholder are replicated to the issuer, who will do the mapping and linking into a combined RDF data store locally (as described in Section 6.2). This combination of the dispersed RDF data sets will take place in a RDF data store that the issuers have to setup and operate within their organization. Parts of this RDF data store will take care of the reasoning over the merged RDF

data set to infer additional triple statements, as well as provide an internal SPARQL endpoint to query the data store from the Web application as shown in Figure 4.8.

Thus, the major tasks will be done on the side of the issuers, which are coordinating and executing the e-commerce fraud investigations as depicted in Figure 6.3.

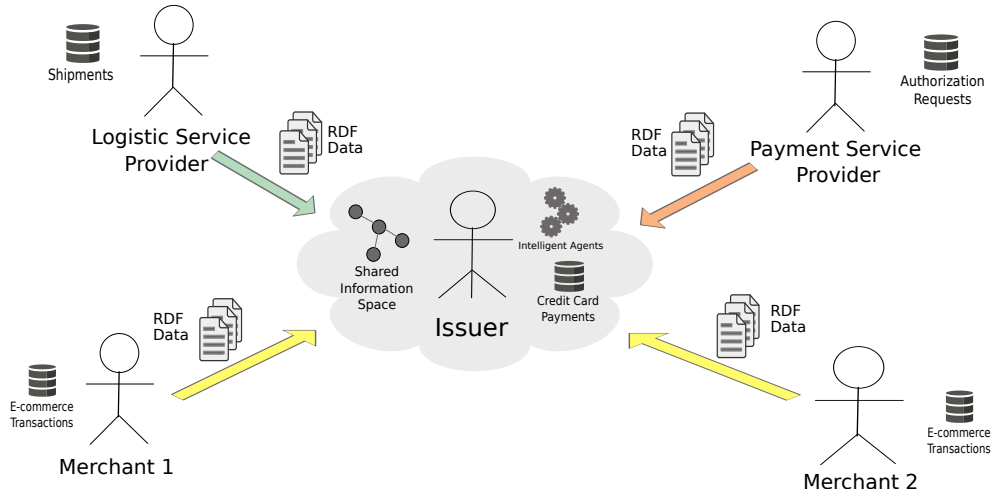


Figure 6.3: Collaborative system using a partially centralized P2P architecture

### 6.3.2 Handling of privacy issues

One of the major concerns with the system architecture mentioned above is that the merchants, PSPs and LSPs have to hand over all of their relevant information to the issuer of a credit card for analysing the e-commerce activities of a consumer. This can raise severe privacy issues, because an issuer will receive a lot of detailed order information from the other parties within this collaborative system. Issuers can not only use these information for validating the correctness of suspicious online transactions, but can also misuse them to build elaborate consumer profiles, which can directly influence the scoring of that consumer in the internal credit and risk rating systems of the issuers.

In addition to that, online merchants will likely not provide to much detail information about their sales and offerings in such a collaborative system, because direct competitors might also be involved in the e-commerce fraud investigation. By sharing parts of the business relevant information, the merchants will raise the fear that their internal business processes, pricing structures, as well as customer loyalty activities get more transparent to their competitors, which might lead to a stronger competition afterwards.

During the design of the collaborative system a valuation of the shared information has to take place, which classify each of them based on the criteria:

- Is the information really necessary to evaluate the e-commerce transactions? Some of the order details from the merchants might not be required to analyse the transactions with the objective to find out about abnormal behaviours.
- Is the information worth protecting? Parts of the necessary information are sensitive information and should be protected against misuse.

Special considerations have to be taken for mandatory *and* sensitive information. In that case, cryptographic algorithms such as hash functions (e.g. SHA-2) can be used to anonymise the information. A hash function is working in one-direction only, and generates a unique hash value based on its input parameter. This hash value is different as soon as the input changes only slightly, and due to the mathematical algorithms used for computing it, a hash value can not be calculated back to the original input value of the hash function afterwards.

A valid use case for such a hashing of information is the e-mail address of a consumer. A plaintext e-mail address such as "max.mustermann@web.de" will always result in a SHA-256 based hash value of "349124ca834949537d726da26dc029e593be72a8b00b81 c47124f5d009c9982b" regardless of the stakeholder, who has originally computed it. That is an additional benefit of the usage of hash functions, because they still allow the information from different stakeholders to be linked together based on their unique hash values.

Another approach to protect sensitive information is to consolidate them in a broader context — a process called data generalization. An example for that would be the item categories, which can be split up into multiple hierarchical layers to narrowly define the affiliations of items to certain product groups. Instead of exchanging item descriptions with the complete set of categories they belong to, merchants can just share items and their top n categories to obfuscate detailed information about the products bought by a consumer. The same approach will also work for location based information, in which the shared information will not contain the exact geographic position from the original data set, but use an approximation to it by stating locations on a broader scope (e.g. district, city or region) instead.

Based on these explanations the information, which is shared in the e-commerce fraud investigation, can be classified into one of the following four quadrants, and should be handled as depicted in Figure 6.4.

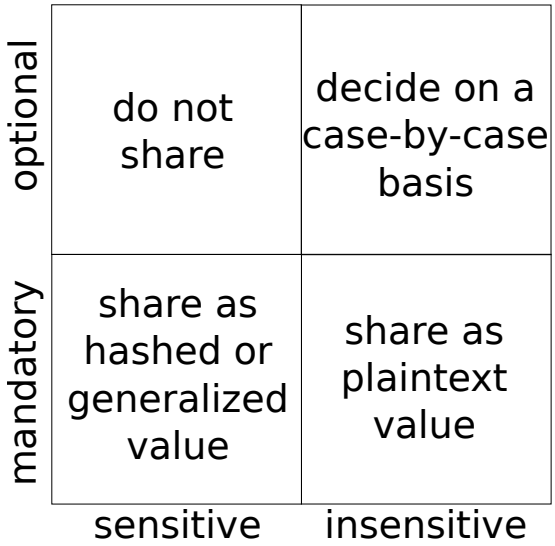


Figure 6.4: Privacy related classification of information in the collaborative system

# 7 Conclusion and Future Work

## 7.1 Conclusion

Initially, this Master Thesis started with three fundamental assumptions:

1. the use of an insecure computer network to transfer credit card information in the e-commerce scenario make them always subject to frauds and counterfeits,
2. the existing technological solutions to detect and prevent them will not be able to cover 100% due to the complexity and dynamics of the whole system, and
3. the substantial analysis of any suspicious transaction is not taken place today due to the enormous effort and time needed to collect all required information manually

As this Master Thesis has pointed out in Chapter 1 and Chapter 3 the first and second premise has been seen a lot of research and development in recent years. But neither an industry-wide usage of the PCI/DSS standards to securely process personal and payment-related information, nor the wide adoption of rule-based or score-based fraud prevention systems has been able to stop deceivers from cheating the e-commerce system and bringing harm to the merchants and consumers alike. Whereas the Master Thesis has shown the impact of e-commerce frauds to the merchants, who are usually the ones that have to cover the losses, it also makes clear that any successful fraud attempt will reduce the thrust into the e-commerce system, which is a substantial part of the global economy already.

Thus, bringing down the amount of fraudulent transactions in the system can not be achieved with technology alone, but require a sharing of information between experts coming from various organizations. As the Master Thesis explained in Section 3.4.3 this consolidation of information about a suspicious transaction is currently not being done in-depth as it would require a lengthy manual communication and analysing process by the investigator of an issuer or PSP. This fact has lead to the third premise, and eventually to the hypothesis that the introduction of a collaborative system, which makes use of Semantic Web and peer-to-peer communication technologies to collect the

information of the relevant stakeholders and link them together, can improve this situation significantly.

As explained in Section 5.4 existing approaches for integrating information from different sources are either to restrictive to work with on a large scale (Web services), or are way to open for the sharing of personal and payment-related information (Semantic Web). Therefore a new proposal for a collaborative system has been developed in Chapter 5 and Chapter 6, which uses suitable aspects from the Semantic Web standards as well as a secure P2P communication network between the stakeholders for collaborating on e-commerce fraud incidents.

Due to the fact that the information will be provided from dispersed organizations, the proposed solution have to deal with differences in structure and wording of them. Under these circumstances the W3C standards such as RDF, RDFS, OWL and SPARQL show their full potential. As part of the Semantic Web initiative they already solve these issues on a global scale. Still, those standards only provide the basics for integrating distributed information. Additional steps are required to build intelligent applications on top of them. In Chapter 5 the Master Thesis showed that the information has to be combined into a graph-oriented representation of the e-commerce activities done with a credit card recently. Based on the initial clustering of order details by merchant a subsequent mapping and linking of the information has to be done, to be able to cluster the transactions on different aspects with the objective to find abnormal activities.

As Chapter 6 explained the Semantic Web standards provide various axioms in the RDFS and OWL specifications to support the mapping of information coming from different sources. Based on these predicates the RDF data store (especially the reasoner in it) can *infer* additional attributes (aka relations) into the combined data set. These mapping expressions can be either managed and injected by the participant, who is doing the merging of the information, or can be integral part of the shared RDF data sets already. Particularly the latter option will ease the integration of different data sources, but requires the participants to look into commonly used RDF vocabularies or ontologies. A selection of the available vocabularies, which might fit the e-commerce fraud investigation, was also depicted in this chapter. In addition to the mapping of the information certain aspects of the transactions have to be linked together to be able to cluster and analyse them in detail. Linking information on the Semantic Web relies on either using the same unique URIs for identical objects in different data sets, or on inferring equality based on property constraints set in the schema definitions.

An explanation of these options can also be found in Chapter 6.

At the end the proposed solution was using a partially centralized P2P architecture that has the issuer of a credit card at its centre. This decision was taken because the issuers are those stakeholders, who will figure out about suspicious activities on a credit card first in the scenario selected for this Master Thesis. They will have to decide whether a transaction is fraudulent or not. In these edge cases, in which the fraud prevention system is unsure about the status of a transaction, an investigator of the issuer will initiate a collaborative session with affected participants. Those parties will share their relevant information with the issuer and collaborate on the investigation of the case. Due to the fact that the issuer as a financial institution has to follow strict regulations and data protection laws, there should be no serious problems for sharing detailed order information with them. Still, some of the information of the stakeholders might be critical and have to be obfuscated in some way. Available methods to do so are part of the proposed collaborative system design.

Thus, this Master Thesis have shown that the technical issues, which lead to fraudulent transactions in the e-commerce scenario, could not be solved completely by introducing new security procedures, but require a collaboration of different experts on each edge case. This collaboration can become more efficient and effective if a shared information space is used, which combine and link together the information from the relevant stakeholders. The existing Semantic Web and P2P communication technologies can act as enablers to design and develop such as collaborative system.

## 7.2 Outlook: Towards a decentralized P2P system

One major concern of the proposed solution can be the partially centralized P2P architecture of the collaborative system, which results in duplicating the relevant information from any stakeholder to the issuer of a credit card in question. Even though the solution offers ways to obfuscate critical information, a *decentralized* P2P architecture for the collaborative system will likely resolve any complains about potential privacy issues. In such an architectural approach each node is equal and keeps their local data. This will have a huge impact on the way the fraud investigation is done. The combined data set will no longer be available for analysing on a single data store with SPARQL, but has to be done in a distributed way as well. Due to this the decentralized P2P system will require new methods for querying, accessing and linking the information. Further research on available protocols and procedures is needed to examine the feasibility of this system architecture.

# List of Figures

# List of Tables

# List of Listings

# Glossary

| | |
|---|---|
| API | Application Programming Interface. |
| B2B | Business-To-Business. |
| B2C | Business-To-Consumer. |
| C2B | Consumer-To-Business. |
| C2C | Consumer-To-Consumer. |
| CSCW | computer-supported cooperative work. |
| CSP | Cloud Service Provider / Hosting Service. |
| e-commerce | Electronic trading over a network such as the Internet. |
| EAN | European Article Number. |
| EMV | Europay, MasterCard and Visa defined security standard for credit and debit cards. |
| ER | Entity-relationship. |
| ETL | Extract-Transform-Load. |
| evt. | eventually. |
| FOAF | Friend-of-a-Friend: commonly used RDF vocabulary to describe people. |
| GTIN | Global Trade Item Number. |
| HTML | Hypertext Markup Language. |
| HTTP | Hypertext Transfer Protocol. |
| ICE | Interactive Connectivity Establishment. |
| incl. | including. |
| IP | Internet Protocol. |
| ISBN | International Standard Book Number. |
| ISP | Internet Service Provider. |
| ISV | Independent Software Vendor. |
| IT | Information Technology. |
| JSON | JavaScript Object Notation. |
| JSON-LD | JavaScript Object Notation for Linked Data. |
| LSP | Logistic Service Provider. |

| | |
|---|---|
| m-commerce | Electronic trading via mobile computers such as smartphones and tablets. |
| NAT | Network Address Translation. |
| OAuth | An open protocol to allow secure authorization on the Web. |
| OWL | Web Ontology Language. |
| P2P | Peer-To-Peer. |
| PCI/DSS | Payment Card Industry Data Security Standards. |
| PSP | Payment Service Provider. |
| RDF | Resource Description Framework. |
| RDFa | Resource Description Framework in Attributes. |
| RDFS | Resource Description Framework Schema. |
| SCTP | Stream Control Transmission Protocol. |
| SEO | Search Engine Optimization. |
| SGML | Standard Generalized Markup Language. |
| SHA-2 | Secure Hash Algorithm 2. |
| SPARQL | SPARQL Protocol and RDF Query Language. |
| SQL | Structured Query Language. |
| TLS | Transport Level Security. |
| URI | Uniform Resource Identifier. |
| URL | Uniform Resource Locator. |
| URN | Uniform Resource Names. |
| vCard | vCard: commonly used RDF vocabulary to describe contact information. |
| W3C | World-Wide Web Consortium. |
| WebRTC | Web Real-Time Communication. |
| XML | Extensible Markup Language. |

# Bibliography

**Allemang & Hendler 2011**
Allemang, Dean; Hendler, James: *Semantic web for the working ontologist: effective modeling in RDFS and OWL*. Elsevier, 2011

**Amazon.com**
Amazon.com: *Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more*. https://www.amazon.com/

**Angiuli et al. 2015**
Angiuli, Olivia; Blitzstein, Joe; Waldo, Jim: How to de-identify your data. In: *Communications of the ACM* 58 (2015), Nr. 12, pages 48–55

**Antoniou & Van Harmelen 2012**
Antoniou, Grigoris; Van Harmelen, Frank: *A semantic web primer*. 3rd. Edition. MIT Press, 2012

**Bannon & Bødker 1997**
Bannon, Liam; Bødker, Susanne: *Constructing common information spaces*. In: *Proceedings of the Fifth European Conference on Computer Supported Cooperative Work* Springer, 1997, pages 81–96

**Bizer et al. 2009**
Bizer, Christian; Heath, Tom; Berners-Lee, Tim: Linked data-the story so far. In: *Semantic Services, Interoperability and Web Applications: Emerging Concepts* (2009), pages 205–227

**Borghoff & Schlichter 2000**
Borghoff, UM; Schlichter, JH: *Computer-Supported Cooperative Work: Introduction to Distributed Applications. Secaucus*. NJ, USA: Springer-Verlag New York, Inc, 2000

**Bouzidi et al. 2014**
Bouzidi, Sabri; Vandic, Damir; Frasincar, Flavius; Kaymak, Uzay: *Product Information Retrieval on the Web: An Empirical Study*. In: *The 8th International Conference on Knowledge Management in Organizations* Springer, 2014, pages 439–450

**Brachmann 2015**
Brachmann, Steve: *In the face of growing e-commerce fraud, many merchants not prepared for holidays - IPWatchdog.com |patents & patent law*. http://www.ipwatchdog.com/2015/11/22/growing-e-commerce-fraud-merchants-not-prepared-for-holidays/id=63271/. Version: 11 2015

**Buford et al. 2009**
Buford, John; Yu, Heather; Lua, Eng K.: *P2P networking and applications*. Morgan Kaufmann, 2009

**Business Wire 2015**
Business Wire: Global card fraud losses reach $16.31 Billion — will exceed $35 Billion in 2020 according to the Nilson report. In: *Business Wire* (2015), 08. http://www.marketwatch.com/story/global-card-fraud-losses-reach-1631-billion-will-exceed-35-billion-in-2020-according-to-the-nilson-report-2015-08-04

**Captain 2015**
Captain, Sean: These are the mobile sites leaking credit card data for up to 500, 000 people A day. In: *Fast Company* (2015), 12. http://www.fastcompany.com/3054411/these-are-the-faulty-apps-leaking-credit-card-data-for-up-to-500000-people-a-day

**Carvalho et al.**
Carvalho, Rodrigo; Goldsmith, Michael; Creese, Sadie; Police, Brazilian F.: Applying Semantic Technologies to Fight Online Banking Fraud.

**Consumer Action 2009**
Consumer Action: Questions and answers about credit card fraud A Q & consumer action A consumer action publication. Version: 2009. http://www.consumer-action.org/downloads/english/Chase_CC_Fraud_Leaders.pdf. 2009. – Forschungsbericht

**DBpedia**
DBpedia: *Online Access*. http://wiki.dbpedia.org/OnlineAccess#1.1%20Public%20SPARQL%20Endpoint

**Dublin Core Metadata Initiative 2012**
Dublin Core Metadata Initiative: *DCMI Metadata Terms*. http://dublincore.org/documents/dcmi-terms. Version: 06 2012

**eBay Inc**
eBay Inc: *eBay: Company Information*. https://www.ebayinc.com/

**García 2006**
García, Roberto: *A Semantic Web Approach to Digital Rights Management*. http://rhizomik.net/html/~roberto/thesis/html/SemanticWeb.html. Version: 2006

**GeoNames**
GeoNames: *GeoNames*. http://www.geonames.org/

**Google Patents**
https://patents.google.com/?q=credit+card+fraud+prevention&after=20150101

**Griffen 2012**
Griffen, Brendan: *The Graph Of A Social Network.* https://griffsgraphs.wordpress.com/2012/07/02/a-facebook-network/. Version: 07 2012

**Grigorik 2013**
Grigorik, Ilya: *High Performance Browser Networking: What every web developer should know about networking and web performance.* " O'Reilly Media, Inc.", 2013

**Grudin 1994**
Grudin, J.: Computer-supported cooperative work: history and focus. In: *Computer* 27 (1994), May, Nr. 5, pages 19–26

**Guha**
Guha, R.V.: *Good Relations and Schema.org.* http://blog.schema.org/2012/11/good-relations-and-schemaorg.html

**Guha et al. 2016**
Guha, RV; Brickley, Dan; Macbeth, Steve: Schema.org: Evolution of structured data on the web. In: *Communications of the ACM* 59 (2016), Nr. 2, pages 44–51

**Hepp 2008**
Hepp, Martin: Goodrelations: An ontology for describing products and services offers on the web. In: *Knowledge Engineering: Practice and Patterns.* Springer, 2008, pages 329–346

**Hoffman et al. 2009**
Hoffman, R. R.; Norman, D. O.; Vagners, J.: "Complex Sociotechnical Joint Cognitive Work Systems"? In: *IEEE Intelligent Systems* 24 (2009), May, Nr. 3, pages 82–c3

**Holmes 2015**
Holmes, Tamara E.: *Credit card fraud and ID theft statistics.* http://www.creditcards.com/credit-card-news/credit-card-security-id-theft-fraud-statistics-1276.php. Version: 09 2015

**Josuttis 2007**
Josuttis, Nicolai M.: *SOA in practice: the art of distributed system design.* " O'Reilly Media, Inc.", 2007

**Koch 2008**
Koch, Michael: *CSCW and enterprise 2.0 - towards an integrated perspective.* In: *BLED 2008 Proceedings*, 2008

**Lewis 2015**
Lewis, Len: *More vulnerable than ever?* https://nrf.com/news/more-vulnerable-ever. Version: 12 2015

**Parameswaran et al. 2001**
Parameswaran, Manoj; Susarla, Anjana; Whinston, Andrew B.: P2P networking: An information-sharing alternative. In: *Computer* (2001), Nr. 7, pages 31–38

**Pienta et al. 2015**
Pienta, Robert; Abello, James; Kahng, Minsuk; Chau, Duen H.: *Scalable graph exploration and visualization: Sensemaking challenges and opportunities*. In: *2015 International Conference on Big Data and Smart Computing (BIGCOMP)* IEEE, 2015, pages 271–278

**Podila 2013**
Podila, Pavan: *HTTP: The Protocol Every Web Developer Must Know - Part 1*. http://code.tutsplus.com/tutorials/http-the-protocol-every-web-developer-must-know-part-1--net-31177. Version: 04 2013

**PYMNTS 2016**
PYMNTS: *Hackers and their fraud attack methods*. http://www.pymnts.com/fraud-prevention/2016/benchmarking-hackers-and-their-attack-methods. Version: 02 2016

**Rampton 2015**
Rampton, John: How online fraud is a growing trend. In: *Forbes* (2015), 04. http://www.forbes.com/sites/johnrampton/2015/04/14/how-online-fraud-is-a-growing-trend/#16ffc0ec349f

**Rana & Baria 2015**
Rana, Priya J.; Baria, Jwalant: A Survey on Fraud Detection Techniques in Ecommerce. In: *International Journal of Computer Applications* 113 (2015), Nr. 14

**Reuters 2015**
Reuters: *Fraud rates on online transactions seen up during holidays: Study*. http://www.reuters.com/article/us-retail-fraud-idUSKCN0T611T20151117?feedType=RSS&feedName=technologyNews. Version: 11 2015

**Rice 1992**
Rice, Ronald E.: Task Analyzability, use of new media, and effectiveness: A multi-site exploration of media richness. In: *Organization Science* 3 (1992), 11, Nr. 4, pages 475–500. http://dx.doi.org/10.1287/orsc.3.4.475. – DOI 10.1287/orsc.3.4.475. – ISSN 1047–7039

**Rietveld et al. 2015**
Rietveld, Laurens; Verborgh, Ruben; Beek, Wouter; Vander Sande, Miel; Schlobach, Stefan: *Linked data-as-a-service: the semantic web redeployed*. In: *European Semantic Web Conference* Springer, 2015, pages 471–487

**Robert & Dennis 2005**
ROBERT, Lionel P.; DENNIS, Alan R.: Paradox of richness: A cognitive model of media choice. In: *Professional Communication, IEEE Transactions on* 48 (2005), Nr. 1, pages 10–21

**Rodrigues & Druschel 2010**
RODRIGUES, Rodrigo; DRUSCHEL, Peter: Peer-to-peer systems. In: *Communications of the ACM* 53 (2010), Nr. 10, pages 72–82

**Schema.org a**
SCHEMA.ORG: *Schema.org Extensions*. http://schema.org/docs/extension.html

**Schema.org b**
SCHEMA.ORG: *Welcome to Schema.org*. http://schema.org

**Sen et al. 2015**
SEN, Pritikana; AHMED, Rustam A.; ISLAM, Md R.: A Study on E-Commerce Security Issues and Solutions. (2015)

**Sobko 2014**
SOBKO, Oleg V.: Fraud in Non-Cash Transactions: Methods, Tendencies and Threats. In: *World Applied Sciences Journal* 29 (2014), Nr. 6, pages 774–778

**Staab & Stuckenschmidt 2006**
STAAB, Steffen (Hrsg.); STUCKENSCHMIDT, Heiner (Hrsg.): *Semantic web and peer-to-peer*. Springer Science + Business Media, 2006. http://dx.doi.org/10.1007/3-540-28347-1. http://dx.doi.org/10.1007/3-540-28347-1. – ISBN 9783540283461

**Sydow 1985**
SYDOW, J: Der soziotechnische Ansatz der Arbeits-und Organisationsgestaltung: Darstellung. In: *Kritik und Weiterentwicklung, Frankfurt/New York* (1985)

**TaskRabbit**
https://www.taskrabbit.com/about

**Taylor & Harrison 2008**
TAYLOR, Ian J.; HARRISON, Andrew: *From P2P and grids to services on the web: evolving distributed communities*. Springer Science & Business Media, 2008

**Virtue 2009**
VIRTUE, Timothy M.: *Payment card industry data security standard handbook*. Wiley Online Library, 2009

**Visa Europe 2014**
VISA EUROPE: *Processing e-commerce payments*. https://www.visaeurope.com/media/images/processing%20e-commerce%20payments%20guide-73-17337.pdf. Version: 08 2014

**Vogt et al. 2013a**
VOGT, Christian; WERNER, Max J.; SCHMIDT, Thomas C.: *Content-centric user networks: WebRTC as a path to name-based publishing.* In: *Network Protocols (ICNP), 2013 21st IEEE International Conference on* IEEE, 2013, pages 1–3

**Vogt et al. 2013b**
VOGT, Christian; WERNER, Max J.; SCHMIDT, Thomas C.: *Leveraging WebRTC for P2P content distribution in web browsers.* In: *Network Protocols (ICNP), 2013 21st IEEE International Conference on* IEEE, 2013, pages 1–2

**W3C 2004**
W3C: *OWL Web Ontology Language Guide.* `https://www.w3.org/TR/2004/REC-owl-guide-20040210/`. Version: 02 2004

**W3C 2013**
W3C: *W3C semantic web activity.* `https://www.w3.org/2001/sw/`. Version: 06 2013

**Wood et al. 2014**
WOOD, David; ZAIDMAN, Marsha; RUTH, Luke; HAUSENBLAS, Michael: *Linked Data.* Manning Publications Co., 2014

# Declaration in lieu of oath

I hereby declare that this Master Thesis was independently composed and authored by myself.

All content and ideas drawn directly or indirectly from external sources are indicated as such. All sources and materials that have been used are referred to in this thesis.

The thesis has not been submitted to any other examining body and has not been published.

Place, date and signature of student
Andreas Gerlach