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 and that 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. Additionally, 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 cloud) can *not* be used.

This Master thesis is therefore analyzing 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: peer-to-peer communication, Semantic Web, CSCW

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, that this thesis will focus 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 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 is usually 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 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 are focusing on 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 usually 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 normally interacting with, but instead navigate them to a malicious web site with the intend 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). Additionally, some of the online retailers are not encrypting 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 in 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 also impact the revenue of the online retailers. Here we have seen a growth of 94% in revenue lost in 2015. Overall it is estimated that credit card fault resulted in $16 billion losses globally in 2014 (PYMNTS 2016) (Business Wire 2015).

While it is possible to prevent fraudulent transactions in the card-present, real-world scenario (mostly due to introducing better technology and establishing organizational countermeasures in the recent 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 fraudulent transactions

to a manageable amount. But this is still an expensive and inefficient solution to integrate into the retailers' business processes, and is largely driven by machine-learning techniques and manual review processes (Brachmann 2015). Additionally, it can be assumed that the online retailers are getting into a "Red Queen race" with the criminals here: with every new technology or method 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 will look into a concept to optimize the collaboration between the affected stakeholders in case of an existing credit card fraud in an E-commerce system. It will *not* look into novel techniques and methods to *prevent* credit card fraud in the E-commerce world. 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, that are also affected by the Internet frauds. In these cases merchants usually try to solve the issues on their own, and getting 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 on the same place) with multiple parties, that 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).

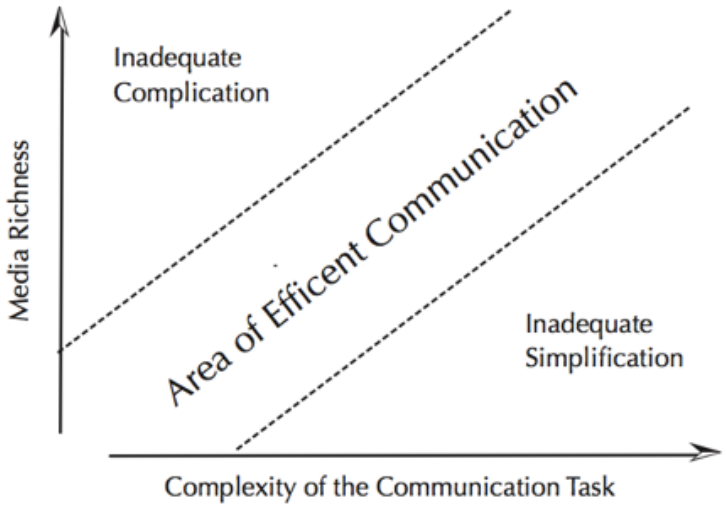

Figure 1.1: The Media Richness Model (Rice 1992)

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 on 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:

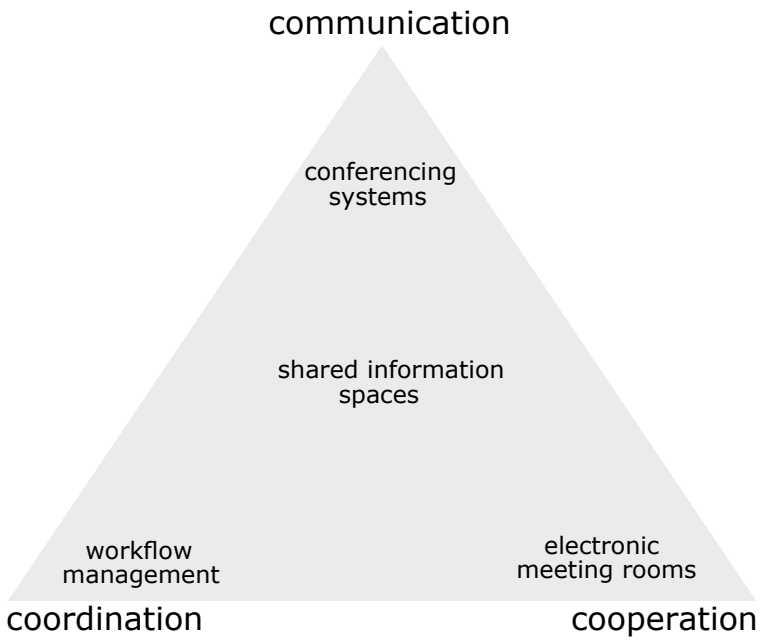

Figure 1.2: The 3C Model (Koch 2008)

A good candidate for such a collaborative system *could* be a shared information space; aka team rooms, cloud storage services or document management systems, that 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 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 Master Thesis Outline

Before starting with the investigation of E-commerce fraud incidents and their possible examinations, the 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 it.

In the next part, Context Analysis in Chapter 3, the 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 thesis briefly outlines the theoretical foundations required for the understanding of the concepts in Chapter 4 and design decisions in 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) com-

munication techniques and protocols.

In the main parts of this thesis (Chapter 5 and Chapter 6) the concept and design for a collaborative system, that supports the investigation of E-commerce fraud incidents, is discussed. These chapters will lay out and analyze 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 at the beginning.

To conclude the thesis also sum up the findings and give an outlook for future work on this topic.

2 Related Works

- “Fraud in Non-Cash Transactions: Methods, Tendencies and Threats.” (Sobko 2014)
- “Overview of E-Commerce” (Ankhule & Joshy 2015)
- “A Survey on Fraud Detection Techniques in Ecommerce” (Rana & Baria 2015)
- “A Study on E-Commerce Security Issues and Solutions” (Sen et al. 2015)
- “Effects of Sensemaking Translucence on Distributed Collaborative Analysis” (Goyal & Fussell)
- “CSCW and enterprise 2.0 - towards an integrated perspective” (Koch 2008)
- “A social network-based system for supporting interactive collaboration in knowledge sharing over peer-to-peer network” (Yang & Chen 2008)
- “Paradox of richness: A cognitive model of media choice” (Robert & Dennis 2005)
- “SWAP: Ontology-based Knowledge Management with Peer-to-Peer Technology.” (Ehrig et al. 2003)
- “RDFPeers: a scalable distributed RDF repository based on a structured peer-to-peer network” (Cai & Frank 2004)
- “P2P networking: An information-sharing alternative” (Parameswaran et al. 2001)
- “Introduction to XMPP protocol and developing online collaboration applications using open source software and libraries” (Ozturk 2010)
- “Peer-to-peer systems” (Rodrigues & Druschel 2010)
- “Leveraging WebRTC for P2P content distribution in web browsers” (Vogt et al. 2013)
- “Let our browsers socialize: Building user-centric content communities on webrtc” (Werner et al. 2014)
- “Taking on WebRTC in an enterprise” (Vogt et al. 2013)
- “High Performance Browser Networking: What every web developer should know about networking and web performance” (Grigorik 2013)
- “Semantic web technologies for the financial domain” (Lara et al. 2007)
- “Security ontology: Simulating threats to corporate assets” (Ekelhart et al. 2006)
- “Applying Semantic Technologies to Fight Online Banking Fraud” (Carvalho et al.)
- “The Semantic Web-Based Collaborative Knowledge Management” (Chao et al. 2012)
- “Open eBusiness Ontology Usage: Investigating Community Implementation of GoodRelations.” (Ashraf et al. 2011)

- “Rule interchange on the web” (Boley et al. 2007)
- “Data linking for the semantic web” (Scharffe et al. 2011)
- “Integrating agents, ontologies, and semantic web services for collaboration on the semantic web” (Stollberg & Strang 2005)
- “GoodRelations Tools and Applications” (Hepp et al. 2009)
- “Drawing Conclusions from Linked Data on the Web: The EYE Reasoner” (Verborgh & De Roo 2015)
- “Schema.org: Evolution of structured data on the web” (Guha et al. 2016)
- “Goodrelations: An ontology for describing products and services offers on the web” (Hepp 2008)
- “A functional semantic web architecture” (Gerber et al. 2008)
- “Towards a financial fraud ontology: A legal modelling approach” (Kingston et al. 2004)
- “Complete query answering over horn ontologies using a triple store” (Zhou et al. 2013)
- “Linked data-the story so far” (Bizer et al. 2009)
- “Linked data-as-a-service: the semantic web redeployed” (Rietveld 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, that 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 catalog 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.

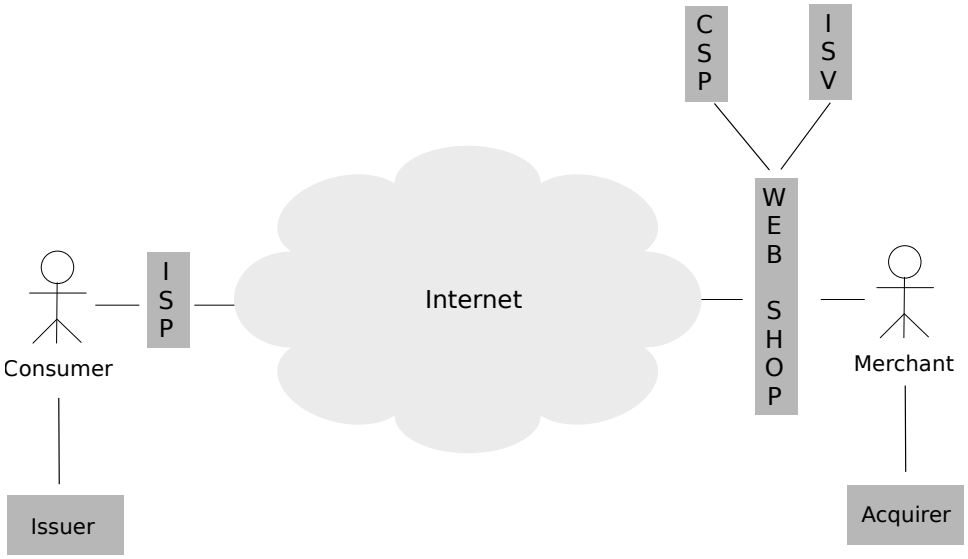

Figure 3.1: E-commerce Fundamentals

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 like 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 is handing over the billing information to the PSP incl. 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 matches 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 an 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), that 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 which 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).

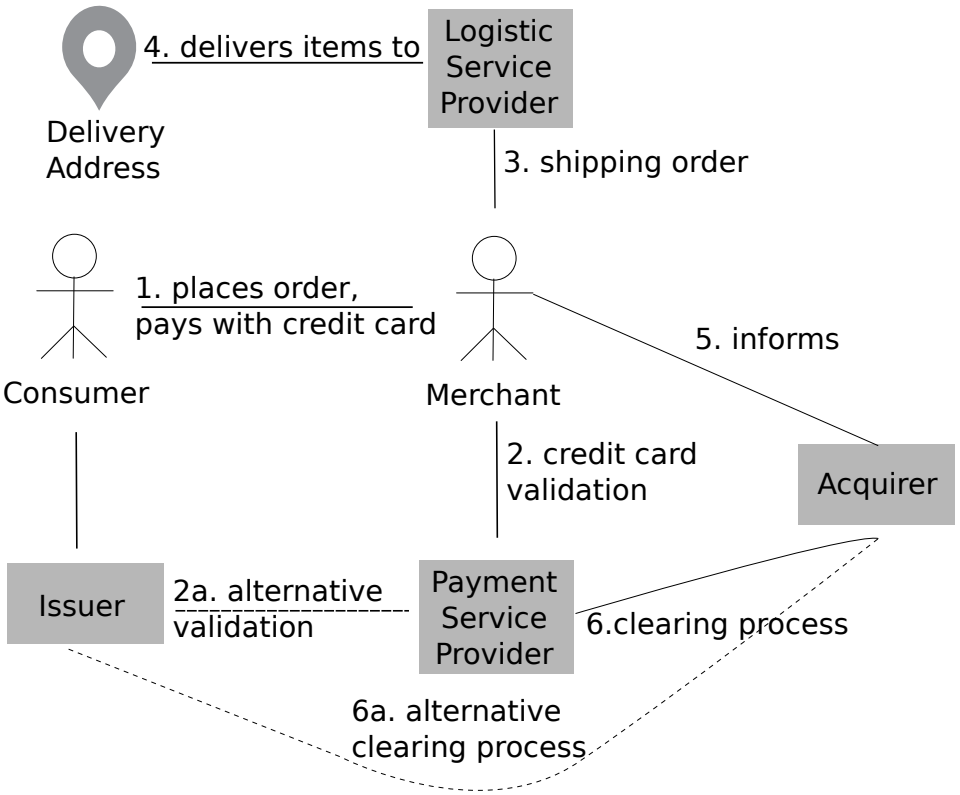

Figure 3.2: E-commerce Checkout Process in detail

3.2.1 Consumer

The consumers are the initiators of E-commerce transactions. They are using 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 handed out from 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 are also having 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 institutions (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 first and put the items of interest into the shopping cart. When finalizing 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 consumer is the main questionable object 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 an URL with a local domain name registry. This specific URL refers to a

fixed public IP address, that 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 hosting or cloud service provider for that. Also 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 allows anonymous users too. The main benefit of the former is the possibility to analyze the shopping behavior 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 if 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 databases. A 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. 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 the consumer's transactions are valid or not. That is due to the fact, that in case of an E-commerce fraud incident the merchants will mostly have to cover the costs (see Section 3.4). Also the online merchant's reputations will suffer, if private information from their databases get 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, that the merchants have to communicate with by sending payment authorization requests to it (see above). The PSPs might be able to authorize a payment request on their own, or might have to route that request to the corresponding issuer of the credit card in question. For the former procedure the PSPs have to run their own databases of registered users with

their credit card information (e.g. a Web service such as PayPal). For the latter they will just have to know, who has issued the credit card in question, 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 are authorizing the payment requests, 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 own back-end 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 holds more of the valuable information to analyze 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 own 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, in whose shops a credit card has been used. Based on public available information from a commercial register about merchants, 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 back-end 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 does usually not process any credit card related information from consumers directly, but refers to the unique payment tokens that have been given out 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 parts of the investigation of fraud incidents as 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 id 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. Still 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. Due to this 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 is found.

3.3 Data flow for credit card transactions

As the previous chapter shows, 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 those will have to deal with the handling of credit card payments and order fulfillments 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.

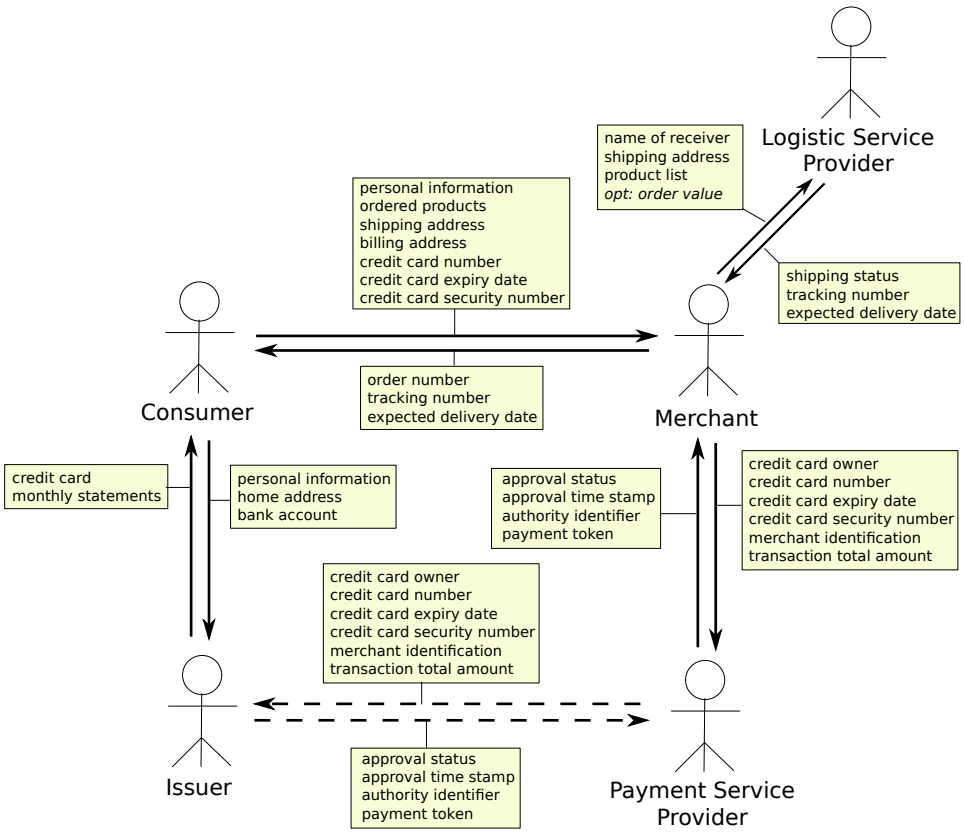

Figure 3.3: Stakeholder and Data Flow in E-commerce scenario

3.4 E-commerce fraud incidents

Based on the previous sections one can come up with strategies fraudsters might use to trick the E-commerce system. To do so the criminals will have to get access to credit card information in the first place. Therefore this section first looks 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 trick 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

Based on information in the Section 3.3 one can figure out the parties, who 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, if a consumer is paying with credit card,
- a Payment Service Provider, if 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, who store and hold the credit card information in their back-end 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). In addition 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. Additionally they will have to send a message to the corresponding issuers, to which the PSPs generally maintain strong business relationships. 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 the checkout processes from the consumers. The credit card information are transfered via the public Internet from the consumers to the merchants and could be victims to man-in-the-middle attacks, in which hackers are intercepting the communication between the consumers and the merchants with the objectives to capture the personal and payment-related information from the data transmission streams. Therefore the merchants should offer their Web shops via a secure communication channel only. For that 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 plain-text information exchanged between consumers and merchants during the checkouts. As the merchants

are not processing the credit card information directly, they also do not have to store them in their own back-end databases. The merchants are asking the PSPs or the issuers of the credit cards for authorization of a payment and receive an 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 abbreviated credit card numbers) 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 the merchants follow these guidelines.

The issuers are a valuable target for hacking into the back-end systems with the objective to leak a massive amount of credit card and personal related information. 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 are doing) 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 ones will 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 are relying 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, that will send them to malicious Web sites with the intend 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-interest. Due to these facts the 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, that was used for its payment, or if 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, that can also be related to as a series of the first fraud activity

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

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 also 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 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 on 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 ca. 70 to 80% of the fraudulent transactions will be recognized as

such and blocked successfully. That still means up to 30 percent of fraudulent transactions could not be identified as such. For handling these edge cases each organization employs special trained staff, that is operating 24/7 and 365 days a year, for managing them.

As stated in the introductory section of this Master thesis, 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 EUR.

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, who 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 the owners of the credit cards, who detect suspicious activities on their credit card accounts and inform their issuers about them. Based on current regulations and laws the issuers have to rollback the fraudulent transactions on request of the consumers, 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 is likely be manageable 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 E-commerce fraud incidents handling

If the fraud prevention systems at the PSPs or issuers are detecting 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, that 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, that 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.

So the operators have to found their decision 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 issuers, who will take over the risk for a false payment authorization. But with an increased value of the items ordered, the PSPs or issuers are putting back the risk to the merchants in case of any consumer complaints later. At a second glance the operators can also verify whether a consumer 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 as they can be bypassed by the fraudsters with ease (see the explanations above).

To make a solid decision the operators will have to get in contact with all the merchants the 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 matches the billing address for that order?
- if not, has the user 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 issuers have had a business relationship with the online merchants in the past. So the operators from the PSPs or issuers might already know whom to contact from the support personnel of the merchants. But in most cases the contact person 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 queries, or there are misunderstandings on handling a case due to language barriers or different incentives between the participants. Additionally the operator, who is responsible for a case, 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 have not been enough, the operator will have 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 affected online merchants. Due to this, 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 fraudulent transactions in detail. Therefore one can assume that an in-depth analysis of any suspicious transaction will not take place today; instead most of these transactions will be acknowledged without any doubt after a first short look and plausibility check.

Still the merchants as well as the PSPs and issuers have a high incentive for increasing the success rates of their fraud prevention mechanisms, and keeping the numbers of successful fraudulent activities low. For the PSPs and 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 institutes to invest in fraud prevention techniques for being able to stay in business. For the merchants it is also of high interest, that a fraudulent transaction can be resolved before a deceiver receives the ordered products. In the worst case scenario of just *one* successfully performed fraudulent transaction in an E-commerce shop this will trigger hundreds if not thousands of subsequent attempts from other fraudsters, as past experiences have shown.

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

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 successful, because there is a lack of information on the side of the merchants as well as the PSPs and issuers. Each of the affected merchants just noticed the transaction that takes place in their own Web shop, without knowing about the other attempts the fraudsters do on the Internet. The PSPs and issuers will both notice the active use of a credit card on different Web shops though, but do not have any detail information about it. Therefore they could not correlate the data from the transactions to check for suspicious activities.

Based on the current credit card usage patterns of the fraudsters, that will try a leaked credit card in commonly used Web shops, it is more likely that these fraudulent transactions will not be recognized on time by the existing fraud prevention techniques in place.

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. These attempts will likely not succeed due to different notions of the communication patterns and data structures 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, that 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**, that 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 thesis though.

4 Theoretical Foundations

This chapter will lay out the theoretical foundations for the to-be-designed collaborative system. It will start with an investigation of the CSCW system theory followed by a detailed examination of the Semantic Web standards like RDF, OWL and SPARQL and how they can be used within Semantic Web agents. Last but not least the chapter will look into the concepts of P2P communication technologies by looking into various protocols for information sharing in detail — e.g. XMPP and WebRTC.

4.1 Computer-Supported Cooperative Work

4.1.1 Definition

4.1.2 Types

CSCW systems can be differentiated by their support of communication on the two axis place and time:

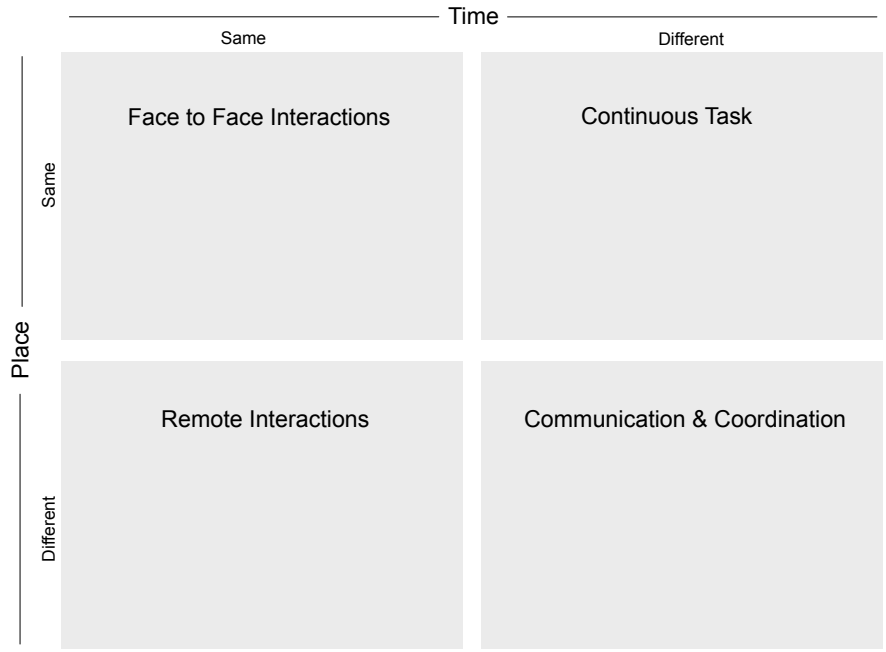

Figure 4.1: CSCW Place/Time Matrix (**?**)

Additionally it is possible to group the CSCW systems based on the 3C model:

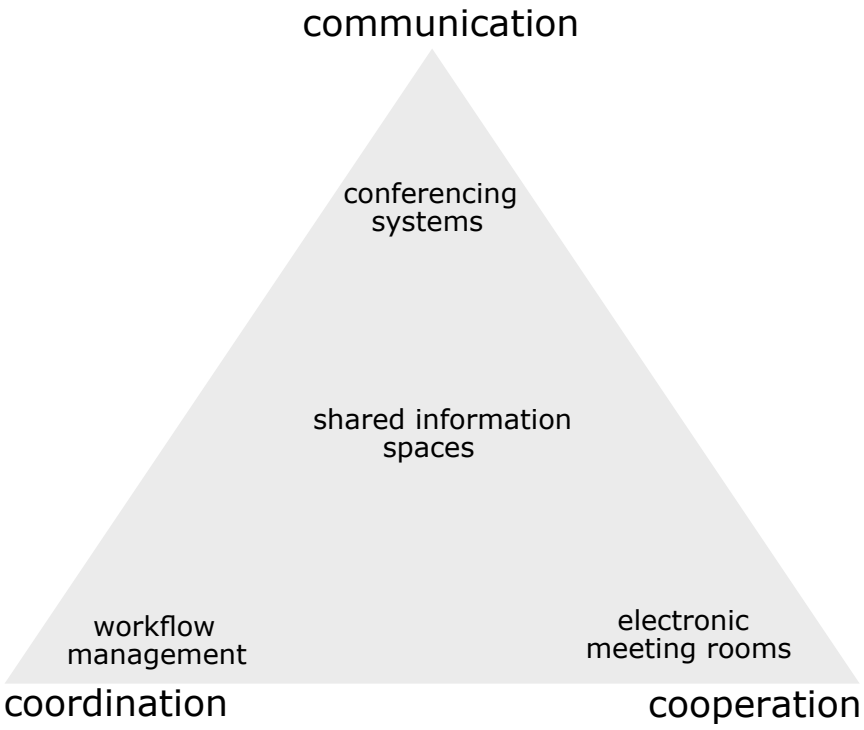

Figure 4.2: The 3C Model (Koch 2008)

4.1.3 Shared Information Spaces

4.1.4 Important aspects of CSCW systems

4.2 Fundamental Web Technologies

4.2.1 The URL concept

4.2.2 The HTTP protocol

4.2.3 The XML format

4.2.4 The JSON format

4.3 The Semantic Web

4.3.1 Vision

MKP Chapter 1:
integrate distributed data from various publishers on the Web into smart applications
the Semantic Web delivers the infrastructure for this vision in form of various standard specifications (RDF, RDFS, OWL, SPARQL, ...)
the fundamentals of the World-Wide Web are also supported by the Semantic Web, especially:
- AAA-Slogan: Anyone can say Anything about Any topic
- Open World Assumption: we must always assume that there exist new information unknown to us yet, that can give additional insights
- Non-unique Naming Assumption: different URIs might refer to the same entity or object

as of this any one can extend on existing data entities and contribute her own knowledge / opinions as well as combine existing information in new ways -¿ data wilderness, no common data schema, more of an organic, living system
it heavily depends on the “network effect” and will / might explode with rising number of users / applications
as there will be disagreements on all sorts of topics there is no single ontology for the whole Web, but rather multiple ontologies tat can be integrated and utilised

MIT Chapter 1:
make information on the Web accessible to machines

- allows integration of information across web sites
- is also known as the “Web of Data”

design principles:
1. make structured and semi-structured data available in standardized formats
2. make individual data elements and their relationships accessible on the Web
3. describe the intended semantics of the data in a machine readable format

HTML is just for human consumption and a lot of the structures and semantics of the underlying databases is lost in the transformation process
- use labeled graphs as data model for objects and their relationships (objects == nodes, edges == relationships between them)
- formalize the syntax of the graph in RDF (Resource Description Framework)
- use URIs to identify individual data items and relations
- use ontologies to represent semantics of the data items (either lightweight RDF schema definitions or Web Ontology Language are used for that)

RDFS and OWL are meta-description languages allowing to define new domain-specific knowledge representations
they rely on the basic principles of the Web: supporting distributed, decentralized architectures

some new initiatives for standardizing semantics: schema.org and linkeddata.org
initially it was tried to solve the integration issues with XML, but as it is syntactically more machine- readable it lacks the semantic of the data
- as of this RDF is the basic language of the Semantic Web and describes meta-data as well as content

an ontology formally describe a domain based on terms and their relationships (terms == classes of objects)
hierarchies are supported (even multiple inheritance between objects)
ontologies also include:
- properties
- value restrictions
- disjointness statements
- specifications of logical relationships
goal is to provide a shared understanding of a domain
can help with the necessity to overcome differences in terminology

a mapping for different wordings in an ontology or between ontologies is possible
they can also be useful for generalization or specialization of Web search results

ontologies help with reasoning of objects, they can uncover unexpected relationships and inconsistencies as well as - by utilizing intelligent web agents - make decisions and select course of actions (e.g. "if-then-conclusions" aka Horn logic)
agents can also be used for "validation of proof" of statements of another agent or machine

Semantic Web is a layered approach ...

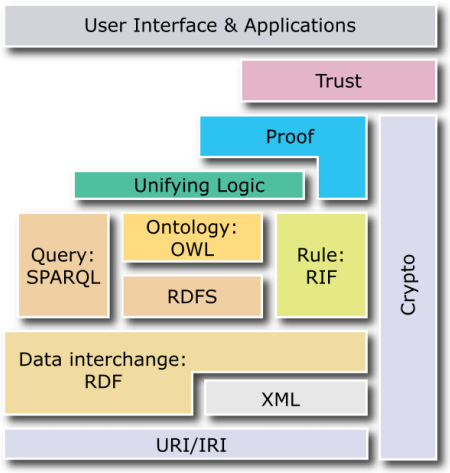

Figure 4.3: The Semantic Web Model (W3C 2013)

4.3.2 Semantic Modelling

MKP Chapter 2:
semantic models
- help people communicate about a fact or situation in the world
- explain and make predictions about the world
- mediate among multiple viewpoints and allow to explore commonalities as well as differences

1. human communication and modelling:
- helps people to coordinate their understanding collaboratively
- knowledge will be gathered, organized, tagged and shared
- when building models in natural human language they are usually open for interpretation of the meaning (e.g. laws)

- interpretation of the text depends on time and context of use -¿ informal model
- the success of informal models can be measured as degree of people supporting the intended purpose
- tagging systems provide an informal organisation to a large body of heterogenous information
- in addition: models can have different layers with an increasing degree of formality (e.g. in the sector of regulations and laws there are regional, national as well as international laws with different degree of formality)
- informal models might be fitting their purpose in the context of their creation, but might need additional layers of models when their usage get beyond that original context to represent the shared meaning

2. explanations and predictions:
- help individuals to draw their own conclusions based on the information received
- especially useful in "interpretive situations" -¿ something is not set in stone
- explanation plays a crucial role in the "understanding" of a situation; if someone can "explain" it, they usually understood it
- in the Semantic Web explanation might help reuse the whole or parts of an existing model
- prediction is closely related to explanation; if a model offer an explanation for a certain situation, it can also be used to make predictions
- that resembles the fundamental of the scientific method (falsification)
- explanation and prediction require a more formal models than used for human communication (see above)
- usually they are build up from objective statements that are used to describe principles and rules (aka formalism)
- these models can also be used to make predictions
- they allow to evaluate the validity of a model and its applicability to a given situation
- in opposite to human communication formalism doesn't need extra layers of explanations
- in the Semantic Web there are certain standards (a formalism) for modelling explanations
- these techniques can also be used to validate proofs and make predictions (aka inference)

3. Mediating Variability:
- goes hand in hand with AAA principle of the Semantic Web
- usually one decides for a specific viewpoint based on the information from trusted

authorities
- informal approach: let every opinion stay side-by-side and let the consumer choose which one to follow
- in this scenario the notion depends on the readers interpretation (as is also common in the Web of information)
- can be modelled in an OOP sense with classes and a hierarchy between them (the higher the more general, the lower the more specific)
- works well for known categories of entities (aka taxonomies)
- any model can also be build up from contributions from multiple sources
- usually seen as layers from different sources
- combination of all layers into a complete model
- a simple merge operation on the layers is easy, but might also introduce inconsistencies of viewpoints into the model
- when two or more viewpoints come together on the Semantic Web there will be an overlap of information
- this will result in disagreements and confusions in the beginning before there will be synergy, cooperation and collaboration
- essence of the Semantic Web: provide an infrastructure that supports AAA and help the community to work through the resulting information chaos to come up with a shared meaning

4. Level of expressivity:
- different people contribute information on different levels of expressivity
- each level might be sufficient to answer specific questions while leaving out unnecessary (sometimes confusing and complex) details
- as of this each level has its purpose!
- also on the Semantic Web there are tools for different levels of expressivity, from the least to the most expressive:
1) RDF: foundation for making statements
2) RDFS: basic notion of classes, hierarchies and relationships
3) RDFS+: subset of OWL, more expressive as RDFS, less complex than OWL, but no standard yet. tries to solve some issues with RDFS for industry use
4) OWL: express logic on the Semantic Web like contraints between classes, entities and relationships

- in the context of the Semantic Web modelling is an ongoing process with some well-structured knowledge and some new, unstructured information coming in at the same point in time

4.3.3 Resource Description Language

MIT Chapter 2:
what is needed to exchange information?
1. syntax: how to serialize the data?
2. data model: how to structure and organize the data?
3. semantics: how to interpret the data?

HTML is made for rendering information on screen and for human consumption
RDF brings a flexible data model to the Web:
- basic building block is a **triple** of *entity* - *attribute* - *value* also known as statement (could also be expressed as *subject* - *predicate* - *object*)
RDFS describes the vocabulary that is available

so:
1. syntax: Turtle, RDFa, RDF-XML or JSON-LD
2. data model: RDF
3. semantics: RDFS

foundational elements are:
- resources (aka just a “thing” of interest identified by an URI or URL depending on its accessibility)
- properties (specify the relations between resources, also identified by URIs)
- statements (assign a value to a ‘resource-property’ relation, value could be another resource or a literal)
- graphs (RDF is a graph-centered data model, could be distributed, Web of Data / Linked Data approaches)

linked data principles:
- use URIs as name for things
- use HTTP URLs so ppl. can look up those things on the Web
- if they do so, provide useful information (HTML and/or RDF, content and/or meta data)
- include links to other URLs so they can discover more/related things

named graph:
- can be used to point to specific statements or (sub-)graphs
- alternative: reification via an auxiliary object

Turtle: Terse RDF triple language
- <subject incl. URI><predicate incl. URI><object incl. URI>.
- literals will be expressed as "value"^^<XML schema data type>and supports *string, integer, decimal, dates, ...*
- URIs can be prefixed: @prefix: <URI>
- repetition: ';' repeats the subject from previous statement, ',' repeats subject and predicate from previous statement
- named graphs in Turtle via Trig extension:
[...] <predicate incl. URI> [...]

```
@prefix ns1: <URI>
@prefix ns2: <URI>
@prefix ns3: <URI>

ns1:subject ns2:predicate ns3:object .
```

RDF/XML: RDF represented in XML format
- RDF namespace and root node
- subjects in 'RDF:description' node containing 'RDF:about' attribute with URI
- predicates and objects are child elements of subject node
- use XML namespeaces for URI of nodes

```
<rdf:Description rdf:about="<subject incl. URI>">
  <ns2:predicate rdf:resource="<object incl. URI>" />
</rdf:Description>
```

RDFa: mixin RDF meta-data into HTML
- 'about' attribute on or in HTML
- 'property' attribute for literal value assignment

- ‘rel’ and ‘resource’ attributes for non-literals
- use XML namespaces for URI of data nodes
- put ‘[]’ around subject and object notations

```
<div about="[ns1:subject]">
  <span rel="ns2:relation" resource="[ns3:object]">
</div>
```

MKP Chapter 3:
- usually data is provided in tables from a database
- if we wanna split those over multiple servers, we can:
1) simply split the tables on a row-basis; the table needs to have the same layout on all servers
2) simply split the tables on a column-basis; the rows in each column need an unique identifier to match up the results
3) break down the whole table into cells and distribute them across all servers
->cells with facts need an unique identifier for the row as well as the column

- therefore RDF uses a triple of subject - predicate - object
- subject and predicate are using an unique identifier based on URI
- the triple can be visualized as directed graph

- data from multiple sources can be combined into a graph, if it can be figured out, which nodes exist in both distributed graphs
- therefore nodes are prefixed with an URI
- this URI should be an URL if the information can be dereferenced on the World-Wide Web
- usually they are used in combination with qnames, which define abbrevations for full-qualified URIs
e.g. qname <URI>
qname:subject predicate qname:object .
- use camel case for identifiers, no spaces are allowed
- W3C defines some qnames themselves:
- rdf: contains identifiers used in RDF
- rdfs: contains identifiers used in RDFS
- owl: contains identifiers used in OWL

- in any case: if you use URLs for your entities at least provide a Web page with the explanation of them

- use rdf:type to specify the type of a subject or object (e.g. geo:Berlin rdf:type geo:City .)
- use rdf:Property to specify an identifier to be used as a predicate (e.g. geo:latitude rdf:type rdf:Property .)
- the references objects could also be literal objects like numbers, dates and strings (they borrow the data type specifications from the XML standard)

- statements can also refer to other statements; this kind of metadata about statements can include:
1) provenance (who has made the statement)
2) likelyhood (what is the probability of this statement)
3) context (the setting in which the statement is valid)
4) timeframe (the time constraints for this statement)

- explicit reification with the predicates rdf:subject, rdf:predicate, rdf:object; e.g.:
q:n1 rdf:subject geo:Berlin
rdf:predicate geo:size
rdf:object geo:MegaCity .

web:Wikipedia m:says q:n1 .
- this sample just qualifies that a source (here: Wikipedia) has made a certain statement (n1); but does say nothing about the statement itself! it is up to the application to decide whether the source (Wikipedia) can be trusted or not!

- RDF triples can be serialized as:
1) N-Triples
2) Turtle
3) RDF/XML
4) RDFa

- blank nodes are commonly used to express unknown or uncertain entities
- they will be described in turtle within []
- an ordered set of items can be represented in turtle as ()

4.3.4 Web Ontologies

Lightweight approach: RDFS
- is about adding semantics to your RDF documents

Start by:
1. specify the **things** to talk about
differentiate between *objects* (real entities) and *classes* (set of entities)
'rdf:type' attribute to assign objects to classes (object = instance of this class)
impose restrictions on the kind of properties used on objects:
- restrictions on values are called 'range' restrictions (object can take values of . . .)
- restrictions on property-object relations are called 'domain' restrictions (this relation applies to objects of . . .)
2. set up relations between classes (inheritance, composition)
3. define properties (registered globally) and the possible hierarchy relationship between them (global properties means you can extend existing RDFS classes with your own properties easily)

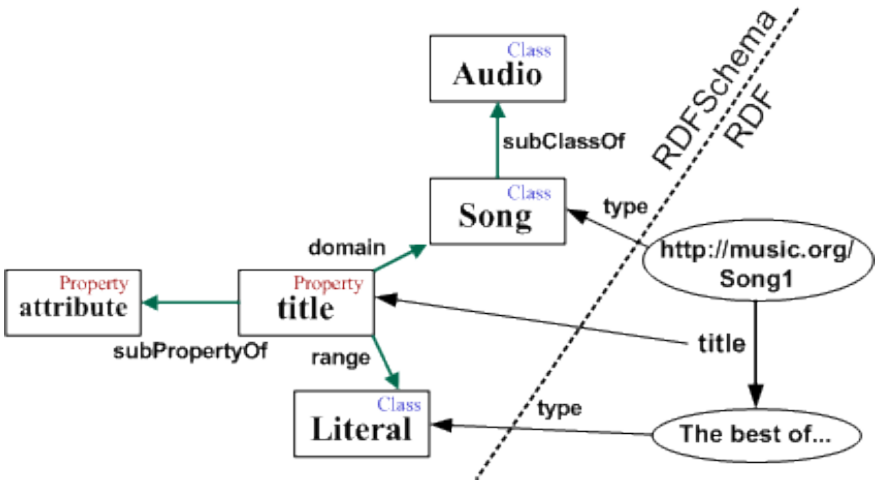

Figure 4.4: RDF Schema sample

RDFS is described in RDF style using:
- core classes like:
- 'rdfs:Resource' (all objects/resources)

- ‘rdfs:Class’ (all classes)
- ‘rdfs:Literal’ (all literals)
- ‘rdfs:Property’ (all properties)
- ‘rdfs:Statement’ (all reified statements)
- core properties like:
- ‘rdfs:type’ (specify kind of class)
- ‘rdfs:subClassOf’ (specify inheritance between classes)
- ‘rdfs:subPropertyOf’ (specify inheritance between properties)
- ‘rdfs:domain’ (specify domain restrictions)
- ‘rdfs:range’ (specify range restrictions)
- container classes like:
- ‘rdf:Bag’ (unordered list of entitites)
- ‘rdf:Seq’ (ordered list of entities)
- ‘rdf:Alt’ (list of alternatives/choices)
- ‘rdf:Container’ (superclass for all containers)
- utility classes like:
- ‘rdfs:seeAlso’, ‘rdfs:isDefinedBy’ (links and references to other entities)
- ‘rdfs:Comment’ (comments and notes of entities)
- ‘rdfs:Label’ (human-friendly name of entities)

Missing features in RDFS: ...

Complex Ontologies in Web Ontology Language (OWL):
...

4.3.5 Query Language

SPARQL requires a **triple store** - a database containing RDF documents
is also referred to as a *Graph Store*
data is inserted via Bulk load operation or via SPARQL update statements
SPARQL consist of SPARQL Queries that are send over the SPARQL protocol
Clients sends the queries to an HTTP endpoint
Stores on the public Web incl. dbpedia.org, ckan.org, wikidata.org
SPARQL also works with RDFS
SPARQL has similarities to SQL: - each element in a triple might be replaced with a

variable like '?varName' like so:

```
PREFIX ns1:<URI>
PREFIX ns2:<URI>
PREFIX ns3:<URI>

SELECT ?varName
WHERE {
    ns1:subject ns2:predicate ?varName
}
```

- in the WHERE clause it hosts the graph pattern to match (could be cascaded to go down subgraphs)
- variables can occur at any place in the graph pattern (?subj ?pred ?obj) as select with query everything

LIMIT <n>option at the end for limiting the result set
FILTER (?varName <condition>) in graph pattern can restrict results to match some literal values and supports:
- numbers, dates: <, >, =
- strings: =, regex()

open world assumption: resources on the Web are described in different schematas with various properties using different vocabularies
- UNION option in graph pattern combines different matches
- OPTIONAL option in graph pattern only returns those entities if they are available (otherwise empty)

ASK query checks for the existence of a given graph pattern
CONSTRUCT can be used to retrieve a subgraph from a larger graph, can also be used to translate between different schemas

```
PREFIX ns1:<URI>
PREFIX ns2:<URI>
PREFIX ns3:<URI>
```

```
4
5  CONSTRUCT {
6      ?varA ns2:predicate ?varB .
7      ?varA ns3:predicate ?literalA .
8  }
9  WHERE {
10     ?varA ns1:predicate ?varB
11 }
12 FILTER ( ?varB > x )
```

- SPARQL can be used to harmonize graphs from different sources
- is also used for basic reasoning ala “if found this, assume that”
- can ease hierarchical queries with * or + on the predicate (SPARQL 1.1)
- can help resolving issues with different entities referring to the same object (MKP pg. 95)
- Federated Queries can be used to combine information from distinct sources via SPARQL (MKP pg. 110-112)

- inferencing information from existing triples via SPIN (SPARQL Inferencing Notation)
- like in a taxonomy items can be categorized in an hierarchy (MKP pg. 114)
- inference patterns are used in Semantic Web applications (MKP pg. 115)
* subClassOf - type propagation rule
- inferencing could be done at query time or persistently (MKP pg. 120/121)
- inferences can also be helpful when combining information from unknown sources
- inferencing happens on various levels (RDFS, RDFS+, OWL) with an increased set of complex inferencing rules (MKP pg. 122/123)

4.3.6 Agents and Rules

4.4 Peer-to-peer communication

4.4.1 Centralized vs. Decentralized Web Architectures

- in a classical client-server scenario a single server is storing information and distributing it to the clients
- the information is centralized and under control of the provider

- a P2P network considers all nodes equal
- each node can provide information to any other node
- information in a P2P network has to be indexed so that the correct node is queried for it
- the index itself has to be stored somewhere (e.g. on a central server like Napster or in a distributed manner spread over the nodes of the P2P network)

- a P2P system has an high degree of decentralization
- the system is usually self-organizing (adding new or removing members automatically)
- the whole system is usually not controlled by a single organisation and spread over various domains
- it tends to be more resilent to faults and attacks
- can be used for file & data sharing, media streaming, telephony, volunteer computing and much more

- can be categorized by the degree of centralization into:
1) partly centralized P2P systems (have a dedicated controller node that maintains the set of participating nodes and controls the system)
2) decentralized P2P systems (there are no dedicated nodes that are critical for the system operation)

4.4.2 Initiating a communication session

- 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 (wellknown IP address)
- in a decentralized P2P system new nodes are expected to obtain via a separate channel the IP address to connect to (usually a bootstrap node that helps to set up the new node)

4.4.3 Finding communication peers

- also known as the overlay network in a P2P system
- can be represented as a directed graph containing the nodes and communication links between them

- can be differentiated between unstructured and structured overlays
- unstructured overlay networks have no constraints for the links between nodes; therefore the network has no particular structure
- structured overlay networks assign an unique identifier from a numeric keyspace to each node; these keys are used to assign certain responsibilities to nodes on the network; as of this routing can be handled more efficiently
- in partly centralized P2P systems the controller is responsible for the overlay formation

- in partly centralized P2P system an object is typically stored at the node that inserted the object
- the central controller holds the information about which objects exist and which nodes hold them

- in unstructured systems the information is typically stored on the nodes that introduces them
- to locate an object a query request is typically broadcasted through the overlay network
- often the scope of the request (e.g. the maximum number of hops from the querying node forward) is limited to reduce the overhead on the system

- in structured systems a distributed index is maintained in the form of a distributed hash table
- this DHT holds the hash value of the (index) key and the address of the node that stores the value

4.4.4 Transmitting Data

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 will discuss the overall concept of such a system on an high level, without going to much into implementation specific details. At the end, the chapter will have answered the question of what the system is and should be able to achieve. In addition to these discussions, the chapter will further look into existing design approaches and analyses why they are of no use for this specific system.

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 have to answer the central question:

Is this transaction really a valid E-commerce transaction?

The relevant stakeholders, that need to be involved in the investigation process, are:

1. **merchants**, who can provide additional information of each E-commerce transaction in question
2. **PSPs/issuers**, that 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 data structures 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 analyze all available information, as depicted in the Figure 5.1.

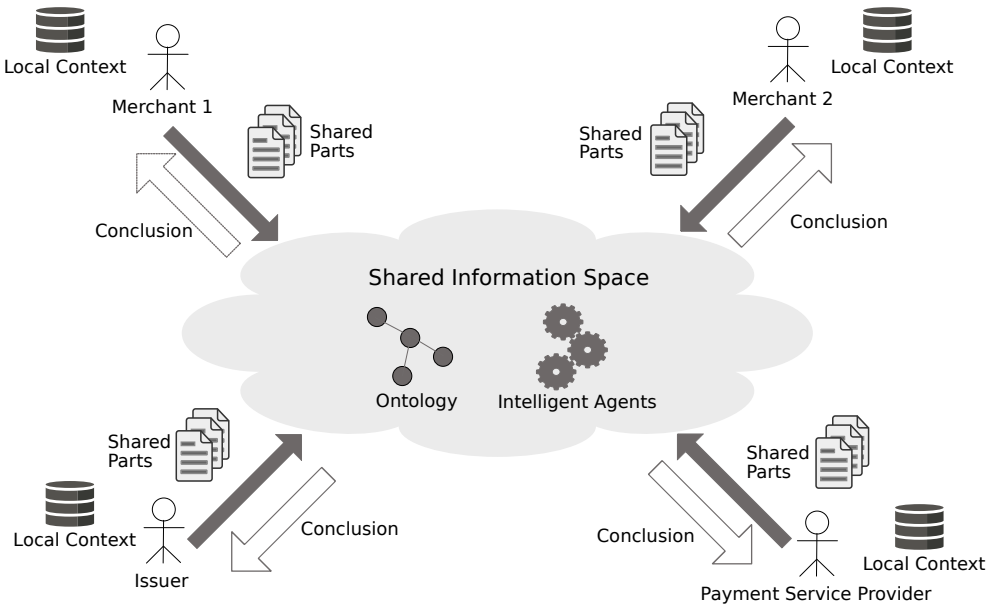

Figure 5.1: High-level concept of the system

In this figure one can see that the relevant parties are providing 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 harmonize and transform the information from each participant into a shared data model to be able to analyze the combined data set. Based on this shared understanding of the E-commerce activities, that have been done with a credit card, a set of intelligent agents (aka analysis algorithms) can assess them and present their findings, that can be valuable to any of the participants of the collaborative system.

5.2 Initial data 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 initial data model can be conducted for the 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

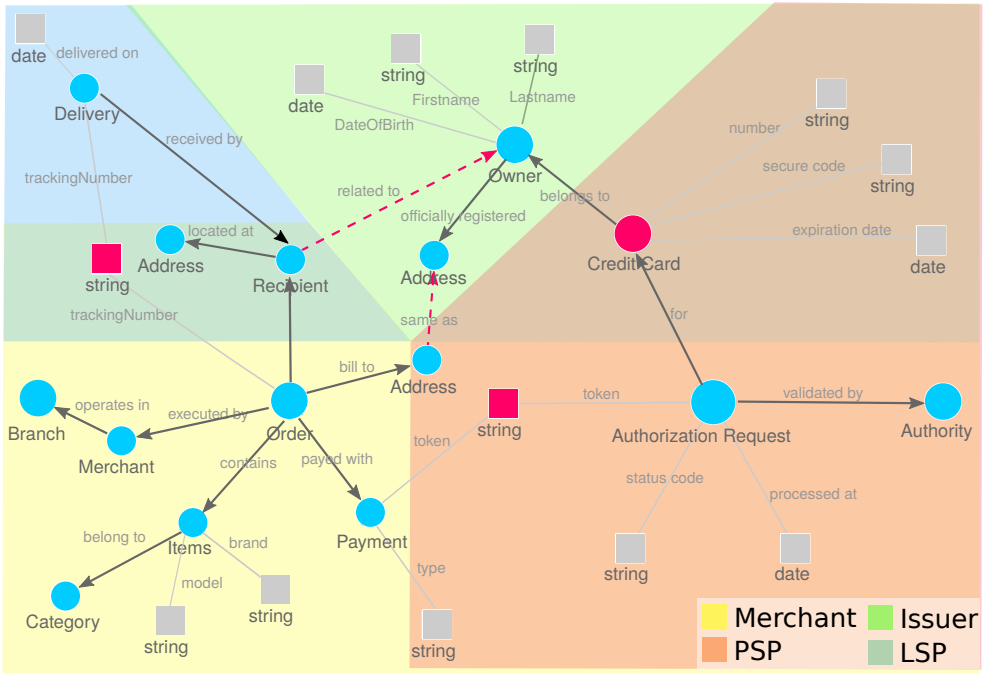

Figure 5.2: Data relations in the E-commerce scenario

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 the Figure 5.2 also shows the important validation criteria. These are connections that have an influence on the decision whether an E-commerce transaction is evaluated as suspicious 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 will have negative impacts on the business success of the online merchants. The PSPs and issuers can not analyze this situation either, as 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 between the relevant stakeholders if the shipping and billing address of an order is different or not 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 the information about the content of the current order. Nevertheless, as mentioned in Section 3.5 looking at the transactions of just one of the merchants is not enough either to solve the E-commerce fraud scenario, that this thesis focuses on. More sophisticated analyzing capabilities are required for the collaborative system to be helpful for the E-commerce fraud investigation.

5.3 Analyzing E-commerce transactions

Therefore the idea is to combine the transaction information from various merchants, LSPs, PSPs and issuers into one combined and shared information space within such a collaborative system to be able to analyze 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. This will also mean that the proposed solution will 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 done with the credit card online recently. For that they will likely have to query the PSPs for the payment tokens first, before asking the affected merchant for order details to any of those payment tokens. At the end each online transaction can be mapped into a schema like the one shown in Figure 5.2, building up a large graph of entities and the relationships between them, and with the specific credit card in the

center of it. An abbreviated sample graph of this procedure can be seen in Figure 5.3.

As shown in this figure the transactions will be clustered by merchants first. Still 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. Especially for those edge cases it might be worth to ask for additional information from the affected merchants to figure out if that 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 analyze the item information (incl. category, brand and model) of each order.

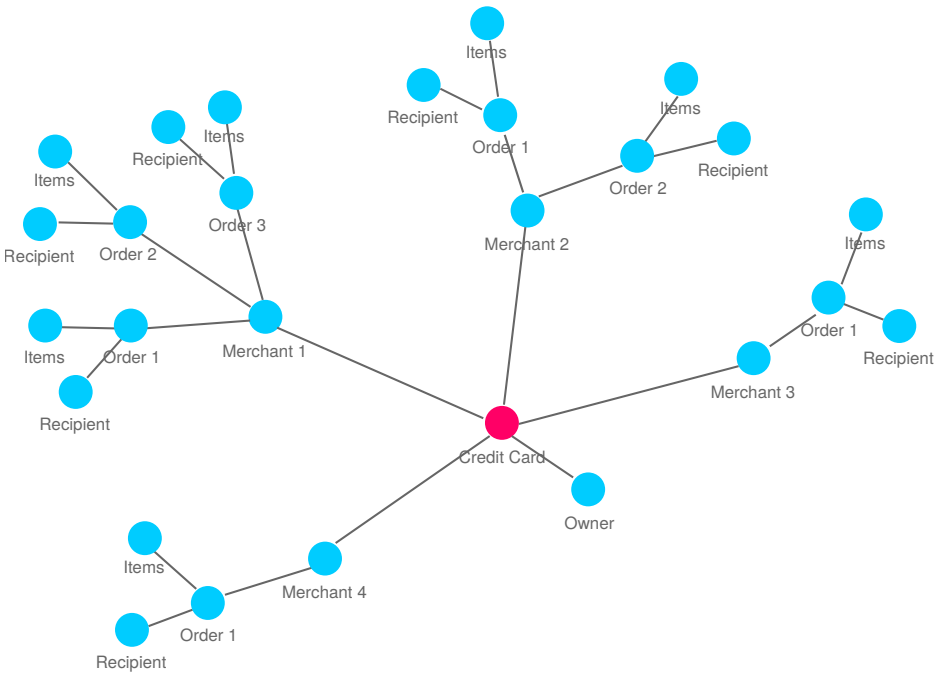

Figure 5.3: Building clusters of E-commerce transactions by merchant

But as already stated, analyzing the cluster of transactions merchant by merchant 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 described in the scenario selected for this Master thesis in Section 3.5. Due to this scenario the various order details from the merchants have to be mapped against each other, so that the initial graph of transactions clustered by merchant can be easily trans-

formed into complementary representations, whose use different criteria to cluster the transactions — such as recipient addresses, branches of merchants, or product-related information. This reshaping of the graph can lead to new insights about the "normal" shopping behavior of a credit card owner, and can make deviations from this behavior visible. Visualizing the combined data set as a clustered graph on screen supports the exploratory nature of knowledge generation and perception, and can therefore help speed up the investigation of E-commerce fraud incidents. An example visualization of a clustered graph, that groups information together based on a criteria, is shown in Figure 5.4. The different colors 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 behavior" can be found in the lower right section of the figure.

Figure 5.4: An example visualization of a clustered graph (Vis.js)

In addition to these clustered graphs the collaborative system can also support the E-commerce fraud investigation by switching the type of visualization based on the criteria chosen for the clustering of the transactions; e.g. when clustering them based on location information such as shipping addresses the system can present the infor-

mation as a heat map on a chart as is displayed in Figure 5.5.

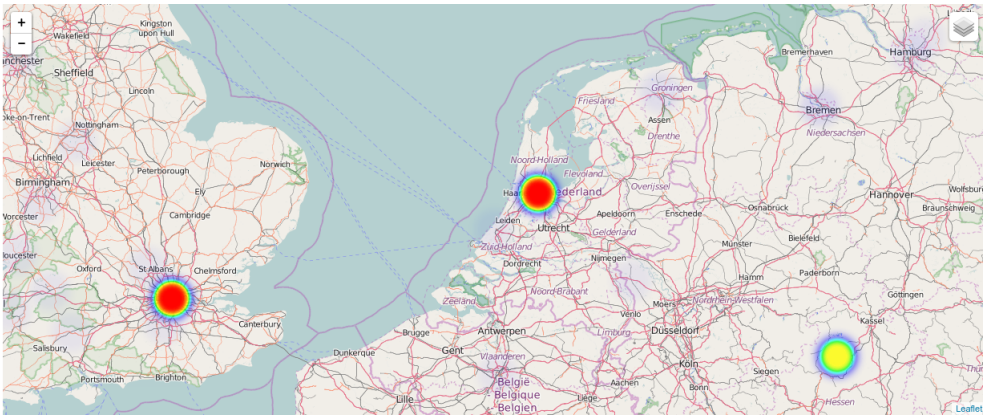

Figure 5.5: Heatmap displaying clusters of location-based information

To conclude the system have to support the collection and combination of E-commerce transaction information from various sources into a large clustered graph, that can be analyzed from multiple view points to validate, if there are any transactions that stand out from the "normal" shopping behavior of the credit card owner. The starting point for the investigation is a sequence of recent credit card activities, that an issuer can provide to the other participants. The graph will initially collect and cluster the information from each merchant based on this list. In case there are suspicious information in one of the transaction clusters of each merchant, 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 of the order detail information between each merchant to allow subsequent analyzing and clustering of the transactions with different criteria.

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 is looking into the common approaches that exist to collect and integrate information between IT systems.

5.4.1 The 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 concept shown in this thesis; namely to get as much information as possible from the various databases, that are in use within a company, harmonize (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 analytics and predictions later. Data within an organization is created and maintained by different business-related tools. Each of these will usually store the information into their own database using a vendor-specific data schema. Other business-relevant data might be stored in structured files, sometimes using a proprietary format such as Excel files. 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 on a separate data store, that holds the combined data set. The whole process is visualized in Figure 5.6.

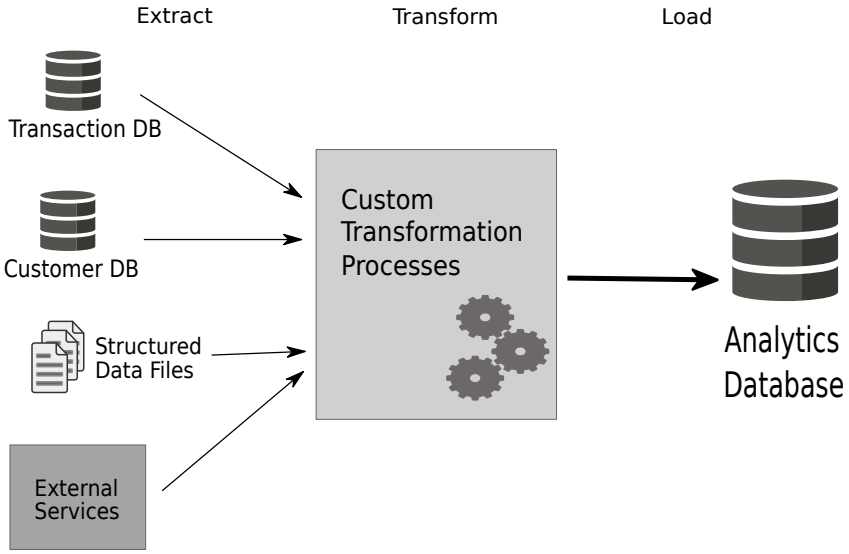

Figure 5.6: ETL process within a company (Wood et al. 2014, pg. 165)

Although this description basically resembles the processes listed in the conceptual section of the E-commerce fraud investigation system, these ETL processes still 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 conditions 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, allowing external partners to access 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. Any changes to one of these would require an elaborate negotiation between the owner of the data source and all of the external partners depending on it.

Beside these drawbacks, that 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. 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" paradigm, that enables 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 language-agnostic. One possible implementation of these concepts are the so-called *Web Services*, that 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, that 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 the 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 check 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; also to distinguish commonly used tag names in an XML document the creator has to place them into specific namespaces (aka XML namespaces). But those XML namespaces further complicate the automatic processing of XML documents and increases 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 usually 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 deserializing the messages, that 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, that want to use the same API from the PSP, can use the same client-side wrapper (sometimes also provided by that PSP for convenience) to send/receive messages to/from this specific Web Service, they still have to make the API-specific integrations into their own Web shops. Also, these integrations are only done in an one-way direction. To allow the merchants to provide information from their own databases, the merchants have to do likewise and provide an API, that others can use to query for information (following the same steps as mentioned above).

Also, as the structures and semantics of the messages and operations of each Web Service interface are not standardized, integrating with other PSPs or issuers result in doing the same integration steps again and again. To make things worse, the mapping of the information coming from different APIs have to be implemented by each client to be able to analyze 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 5.7.

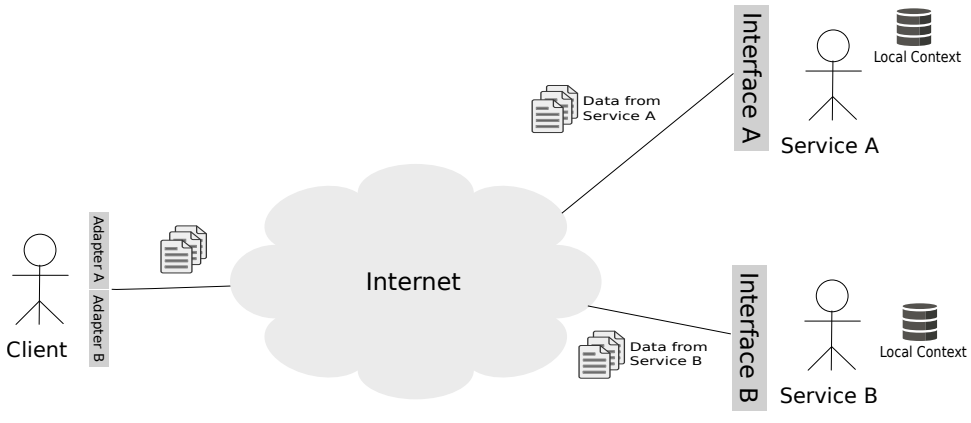

Figure 5.7: Data integration within the Web Service approach

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 retailers, PSPs and issuers as participants, and therefore left out smaller companies from the E-commerce fraud investigation process. For a solution of the problem described in Section 3.5 this is not sufficient. As of this 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 driven by the data available to them. Information that are likely coming from different places in the global information space — accessible usually via a custom API on a server hosting those resources (see Section 5.4.2). The more consistent the information available to the smart Web application is, the better the Web service and its result will be. But to support an integration of the data from various Web services the semantics of the information delivered by each service has to be available — and there has to be a generalized, formalized way to express the semantic of that data. The focus on a standard, that allows Web services to express the semantics of the data, also allows for global scalability, openness and decentralization, which 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 like 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 semantic (aka meaning) 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 an unique URI. These URIs can be abbreviated via "prefix" definitions to make the whole statement easier to read (see also Section 4.3). To specify that there is an order "12345" from a "merchant1", one can come up with the following RDF statement, that uses the Schema.org specification (Schema.org b) to describe an order:

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

Listing 1: An order specification in RDF

An RDF file 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 the above Listing 1 can be visualized as graph like this (see Figure 5.8):

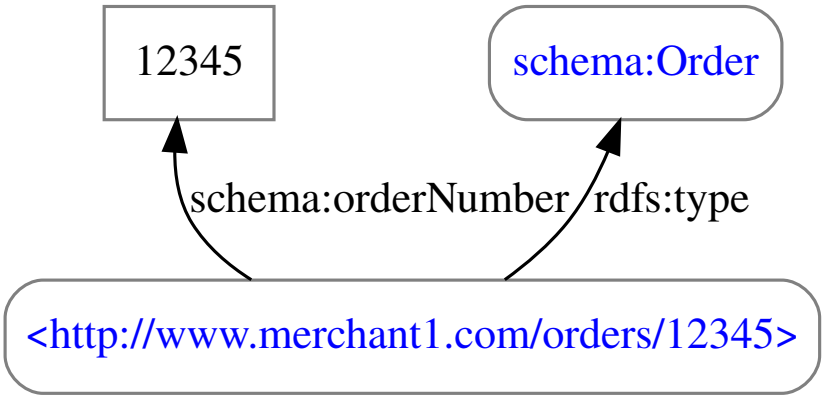

Figure 5.8: Graph-based visualization of the order from Listing 1

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 case merging the "triples" from different RDF data files will result in a local graph holding the combined information as shown in Figure 5.9.

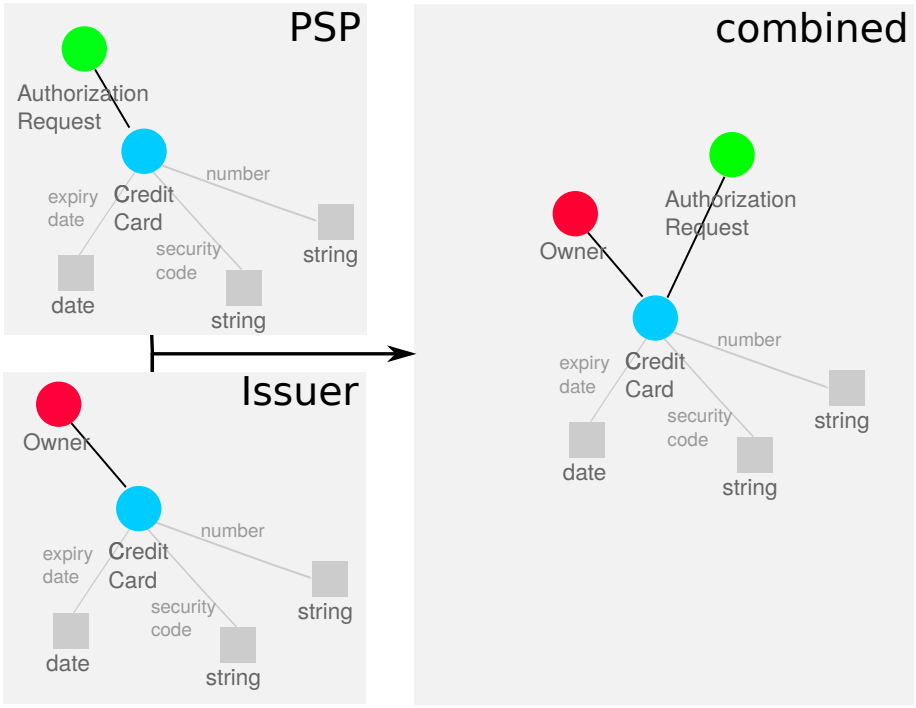

Figure 5.9: Combining two RDF files containing the same credit card entity

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 lying 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 organisation 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 the 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 SPARQL protocol and query language, see Figure 5.10.

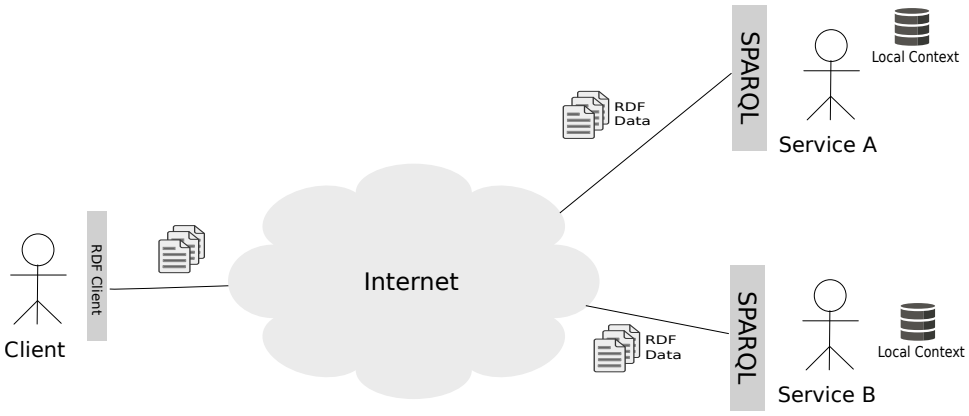

Figure 5.10: Data integration within the Semantic Web approach

As the underlying model of a RDF data set is resembling a graph-based representation it will fit the concept of the proposed system from Section 5.3 perfectly. Still requiring any participant to setup and operate a public 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, requiring 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, who 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 2). 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), that is offering publicly available information from Wikipedia articles in RDF format.

```
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
4 SELECT ?s ?p ?o    # select every subject-object-predicate triple found
5 WHERE {
6  ?s ?p ?o          # do not specifying a condition returns everything
7  }
```

Listing 2: Retrieving all information in an RDF store using SPARQL

To conclude, 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 proposed solution.

5.5 Conclusion

As the previous section showed, existing approaches are of limited use for the design of a collaborative system to support the E-commerce fraud investigation scenario described in Section 3.5. The leading approach for such a 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 parts of information can be returned, and execution of operations can be restricted

Looking at the Semantic Web approach, it's most interesting functionalities are:

- providing information in a semantically self-contained way
- the ability to merge information from different RDF data stores locally
- the graph-based data model underlying the RDF data stores
- the usage of SPARQL to query and analyze the locally combined data sets

In the following Chapter 6 the 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 from the relevant participants of the E-commerce fraud investigation scenario.

6 Design of a collaborative system

This chapter about the design of a collaborative system to support the investigation of E-commerce fraud incidents will start 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 to the relevant parties based on the E-commerce fraud investigation scenario described in Chapter 3. For this purpose it will have a detailed look into the partially centralized P2P communication architecture and show how that can be used within the system for securely sharing the relevant information.

6.1 Choosing a RDF schema

As a major objective of the E-commerce fraud investigation system is to bring the various transactional information from online merchants, PSPs and issuers together, combine them and analyze the resulting graph from different view points, the information exchanged between the relevant participants either have to follow a common schema or have to be mapped against each other.

6.1.1 Re-using vocabularies available on the Web

One valid approach to come up with a data schema is to take a look into commonly used RDF schemata and vocabularies, 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 this 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 schema specifications describing a fictive consumer named “Max Mustermann” incl. his home address can be done by combining data utilizing the FOAF and vCard namespaces in a RDF data set, such as described in Listing 3 and visualized as graph in Figure 6.1.

```
@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 3: Personal related information about a fictive consumer in RDF

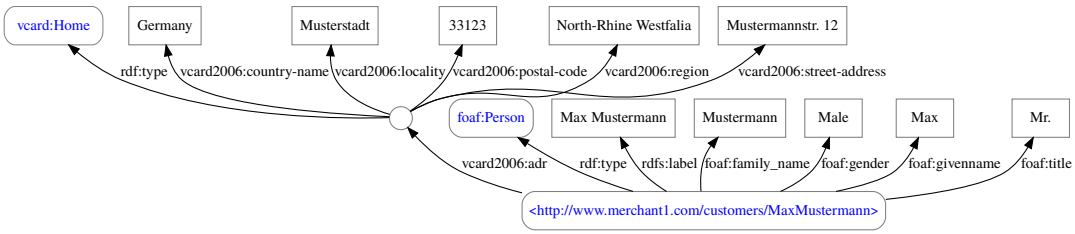

Figure 6.1: Graph representation of consumer information from Listing 3

Looking back to the initial data model from Section 5.2 one can map the information, 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 data 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). Additionally some of the vocabularies are no longer actively maintained, such as GoodRelations. Due to these circumstances one usually have to build an own ontology that fills in the missing pieces and refers to the existing concepts whenever appropriate.

6.1.2 Creating a vocabulary for E-commerce transactions

Another possible approach is to define a completely new schema for the proposed system and share that with every possible stakeholder. This schema will define all the entities and relations known to the collaborative system and would be expressed in RDFS format.

Trying to model the information of a credit card as displayed in Figure 5.2 will result in the RDFS specification shown in Listing 4. This definition of a credit card resource explicitly reuses specifications from the FOAF and GoodRelations ontologies by specifying 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

```
@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 4: A specification for a credit card in RDFS

As most of the parts of the E-commerce data model shown in Figure 5.2 can not be expressed directly with the existing RDF vocabularies, filling in the gaps would mean to come up with a large set of custom entities and relationships. A major drawback of this approach is, that new partners of the system will first have to implement the conversion of their internal data structures to an RDF data set, that is compatible with the specific schema definition, before being able to participate in it. This will limit the general usage of the collaborative system.

6.1.3 Schema.org initiative

When looking back at the list of existing ontologies and vocabularies, that are actively used on the Web today, one will also find the Schema.org vocabulary definition (Schema.org b). This 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, 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 specification (Schema.org a). In one of the past releases of the Schema.org core specification the maintainers also included all of the existing concepts of the GoodRelation ontology into the Schema.org vocabulary (R.V. Guha).

As the merchants will likely provide semantic meta data for their products to improve their listings on search engine results (also known as SEO) using the vocabulary of Schema.org already, one can re-use parts of these information for the E-commerce fraud investigation scenario. Additionally, the wide-ranging scope of aspects declared in the Schema.org vocabulary, make it a good fit for the collaborative system of the E-commerce fraud investigation scenario as one can assume, that each participant is aware of this meta data initiative, and the relevant communication partners will have a common understanding of the terms declared in that system. When trying to map the initial data model from Section 5.2 to the Schema.org core specification, one will basically come up with a schema as displayed in Figure 6.2.

6.2 Working with RDF data sets

6.2.1 Preparing information for external usage

...

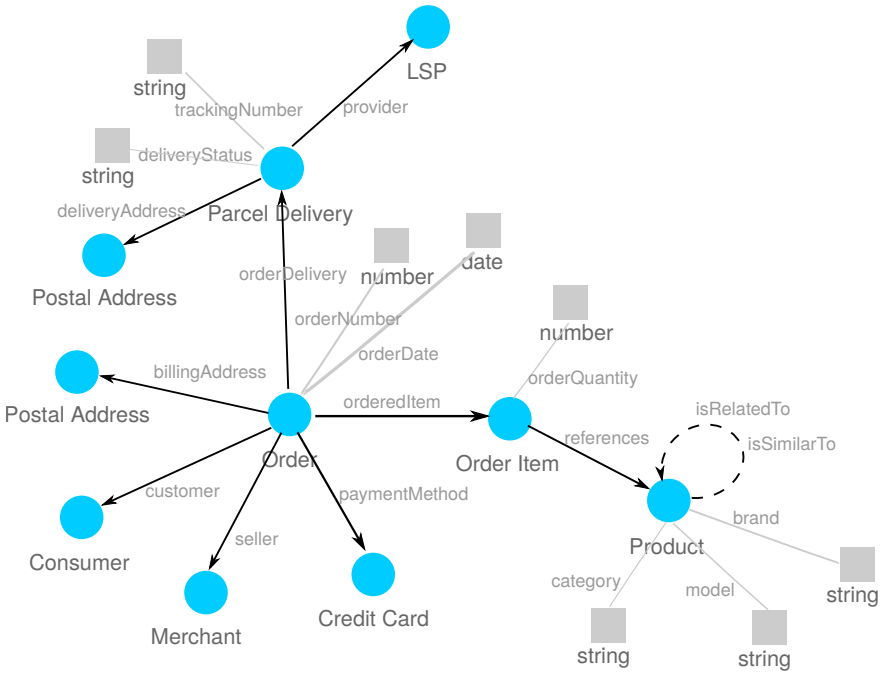

Figure 6.2: Schema.org based mapping of an E-commerce transaction

6.2.2 Mapping of the information from various sources

Although it is possible to model each E-commerce transaction with the Schema.org specification as shown in Figure 6.2, the collaborative system still has to take care of the mapping of the transactional information coming from various sources to be able to combine, analyze and cluster them. As shown in the Section 5.4.3 the build-in merging capabilities of the RDF specification rely on the URIs used for the entities in the various RDF data sets. These URIs are used to uniquely identify the resources in a RDF data set.

The W3C standards for the Semantic Web also include support for these mapping issues, as they will also come up when trying to combine semantic information available around the Web. Additionally, the Schema.org specification also defines a property, that can be used for that purpose. The following properties are available in the RDFS, OWL and Schema.org specifications:

- **rdfs:label**: a label of a resource in the RDF data set can contain a human-friendly name of the resource. These labels are literals of type string and can come with a language specifier in case the resource supports expressions for different languages (see Listing 5 for an example).

- **rdfs:seeAlso**: a relation of type seeAlso contains an URI to an external resource, that contains additional information for the subject (see Listing 6 for an example).
- **rdfs:isDefinedBy**: the isDefinedBy property is a specialization of the seeAlso relation, in that it specifies a link to the original definition of a resource
- **owl:sameAs, schema:sameAs**: the sameAs relation as specified in the OWL and Schema.org vocabularies are providing an unique URI that unambiquiously define the subject (see Listing 7 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 xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix schema: <http://schema.org/> .
@base <http://www.merchant1.com/catalog/> .

<P12345>  rdfs: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 5: Specifying a product with labels in three different languages in RDF[1]

[1]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.merchant1.com/customers/> .

<MaxMustermann> rdf:type schema:Person;
                rdfs:label "Max Mustermann"@en;
                schema:adress [
                  rdf:type schema:PostalAddress;
                  schema:addressLocality "Cologne"@en;
                  rdfs:seeAlso <http://sws.geonames.org/2886242/>
                ] .
```

Listing 6: Specifying a link to a Geo Names resource for looking up additional location information in RDF[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 schema: <http://schema.org/> .
@base <http://www.merchant1.com/customers/> .

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

Listing 7: Specifying a link to a DBpedia resource to uniquely identify an entity in RDF

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

When looking at the E-commerce transaction schema as defined in Section 5.2 the following information must be uniquely identified and mapped within the RDF data sets coming from the participants of the collaborative system:

- **personal related information** such as the consumer, recipient and credit card owner
- **location based information** such as the billing and shipping address as well as the location a credit card owner is registered for
- **product related information** such as the categories, subcategories, brand, model and item description
- **merchant related information** such as the branch of a merchant

To support the unique identification of entities in the RDF data set of an E-commerce transaction, one can refer to publicly available RDF data sets on the Internet, such as GeoNames or DBpedia. These can provide an unique URI for locations and named places. A product-related RDF data set was available in form of the ProductDB initiative until recently (Bouzidi et al. 2014). Due to the shutdown of it, the mapping of products can no longer be done by referencing unique URIs from the Web, but will have to be based on the global trade item number (aka GTIN) of each product. Additional aspects of an item, such as brand, categories and subcategories, can be found on DBpedia as well.

A problem, that will come up, is the unique addressing of personal related information, such as identifying the consumer. The collaborative system can not rely on mapping the personal related information based on properties such as familyName and givenName alone (see Listing 3). There could be typos in the information coming from various RDF data sets, and different individuals can still have the same name information. One possible approach to bring these information together would be the mapping based on the e-mail address of the individual. An e-mail address like an URI is a globally unique addressing scheme, and one can assume that two entities, who are using to the same e-mail address, are referring to the same entity. Still this is only a weak hint as an individual can have more than one e-mail address, and could use different e-mail addresses for the online shopping trips at different merchants. Therefore a more sophisticated mapping algorithm for personal related information is needed in the collaborative system. This algorithm may take into account the combination of familyName, givenName, dateOfBirth as well as location-based information to uniquely

identify an individual. To sum up, the 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	eventually e-Mail address	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

6.3 Building a partially centralized P2P system

For the E-commerce fraud scenario, that has been selected for this thesis in Section 3.5, one can say that the issuer of a credit card is the party who initiates the collaborative fraud investigation. They are recognizing the active use (and likely misuse) of a credit card in the online and the offline world first, and are also getting a notification about any suspicious transactions made with it from their fraud prevention systems. Due to this fact, one can come up with a partially centralized P2P architecture for the E-commerce fraud investigation system, in that the issuer of a card in at the center and acts as a trusted party in this system. This issuer will initiate a collaborative session with the other required stakeholders based on the usage history of the credit card in question. During this P2P communication session the merchants, PSPs and LSPs will share the required information with the issuer. In this process the data from the other stakeholders will be replicated to the issuer, who will build up a networked graph based on the Schema.org specification. So the main work will be on the issuer's side, who is the major driving party in the system, as depicted in Figure 6.3.

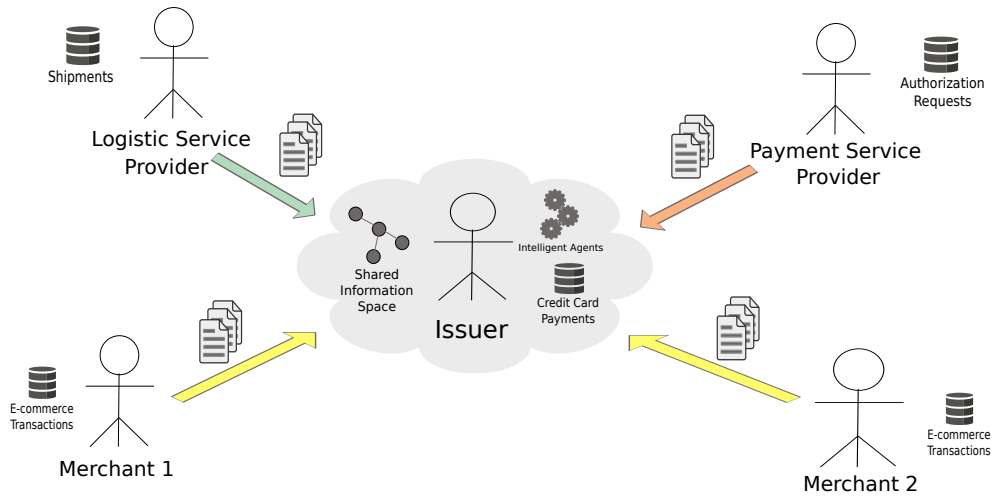

Figure 6.3: Collaborative system using a partially centralized P2P architecture

Using the WebRTC communication protocol for initiating the P2P session will allow the issuers to setup a communication between the relevant stakeholders directly from within an application running in their Web browsers. The application can visualize the connectivity status of the participants, the progress of their data sharing efforts as well as offer direct face-to-face communication possibilities in case of misunderstandings or further requests. A wireframe of the Web application screen is depicted in Figure 6.4.

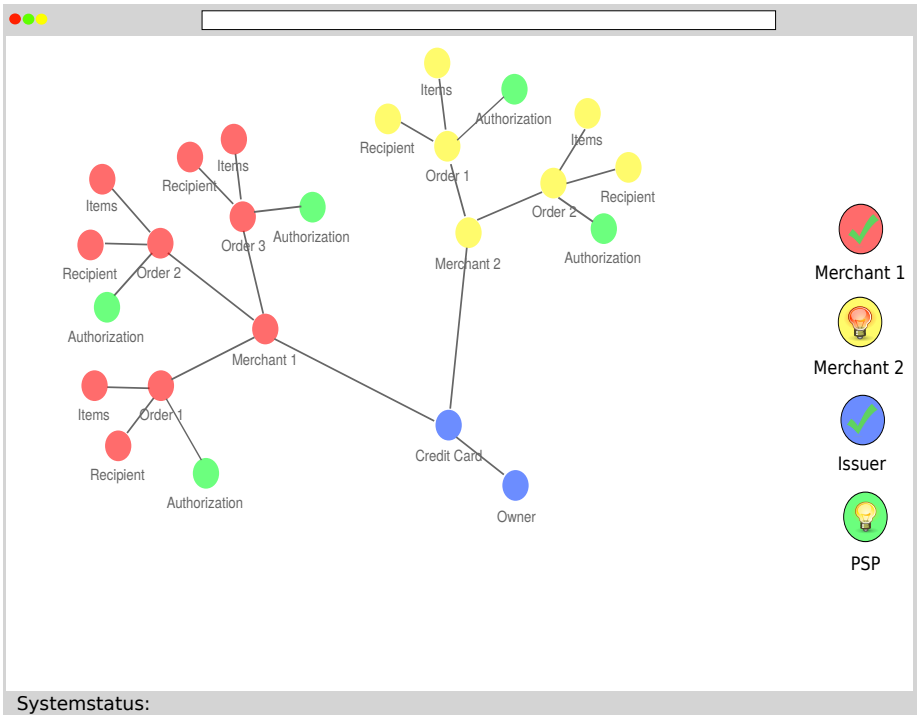

Figure 6.4: Screen prototype of collaborative system showing participants and shared information

One of the major issues with the above mentioned system architecture is, that the merchants, PSPs and LSPs have to hand over all of their relevant information to the issuer of a credit card for the analysis.

...

7 Conclusion and Future Work

7.1 Towards a decentralized P2P system

In the decentralized P2P system architecture each node is equal and keeps their local data ready for analysis if the node is online. If the issuer will have to figure out, whether a transaction is fraudulent or not, she is going to send out various queries to all the available nodes in the P2P cluster asking for certain information that help investigating the case. The other nodes, whose reside on each stakeholder involved, will answering the queries based on the common Schema.org data mapping shown above and send back the results to the issuer bank. The issuer will collect all the results from the various parties and combine them to be able to analyze the issue and come up with a conclusion. The main benefit of this architecture is, that there is no need to duplicate the data from the other stakeholders to the issuer. Due to this it can also be a better suited solution if data sharing faces restrictions due to law or regulations. On the other hand this architecture will depend on the nodes being online all the time so the issuer can query for information at any time. So this works only in synchronous communication mode. Additionally there are efforts spread around all the stakeholders to set up and maintain a system for secure data querying functionality, please see Figure 7.1.

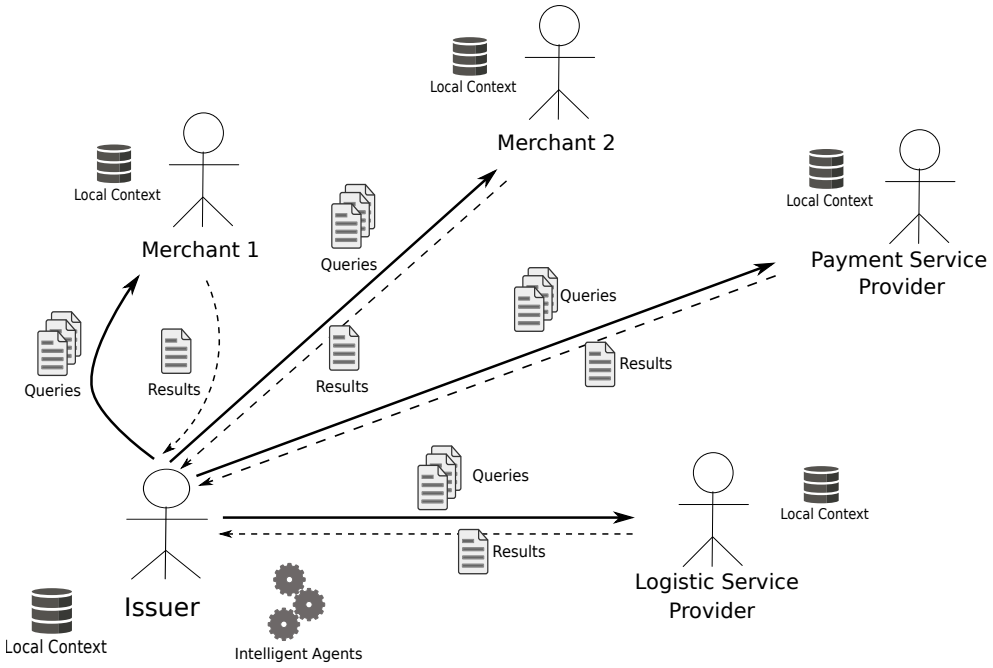

Figure 7.1: Decentralized P2P system architecture

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.
EMV	Europay, MasterCard and Visa defined security standard for credit and debit cards.
ETL	Extract-Transform-Load.
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.
IP	Internet Protocol.
ISBN	International Standard Book Number.
ISP	Internet Service Provider.
ISV	Independent Software Vendor.
IT	Information Technology.
JSON	JavaScript Object Notation.
LSP	Logistic Service Provider.
M-commerce	Electronic trading via mobile computers such as smartphones and tablets.
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.
RDFS	Resource Description Framework Schema.
SEO	Search Engine Optimization.
SGML	Standard Generalized Markup Language.
SPARQL	SPARQL Protocol and RDF Query Language.
TLS	Transport Level Security.
URI	Uniform Resource Identifier.
URL	Uniform Resource Locator.
vCard	vCard: commonly used RDF vocabulary to describe contact information.
W3C	World-Wide Web Consortium.
WebRTC	Web Real-Time Communication.
XML	Extensible Markup Language.
XMPP	Extensible Messaging and Presence Protocol.

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/`

Ankhule & Joshy 2015
Ankhule, Gayatri R.; Joshy, MR: Overview of E-Commerce. In: *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)* (2015), pages 196

Ashraf et al. 2011
Ashraf, Jamshaid; Cyganiak, Richard; O'Riain, Seán; Hadzic, Maja: *Open eBusiness Ontology Usage: Investigating Community Implementation of GoodRelations.* In: *LDOW*, 2011

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

Boley et al. 2007
Boley, Harold; Kifer, Michael; Pătrânjan, Paula-Lavinia; Polleres, Axel: Rule interchange on the web. In: *Reasoning Web.* Springer, 2007, pages 269–309

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

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-accor

Cai & Frank 2004
Cai, Min; Frank, Martin: *RDFPeers: a scalable distributed RDF repository based on a structured peer-to-peer network.* In: *Proceedings of the 13th international conference on World Wide Web* ACM, 2004, pages 650–657

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.

Chao et al. 2012
Chao, Lemen; Xing, Chunxiao; Zhang, Yong: *The Semantic Web-Based Collaborative Knowledge Management.* INTECH Open Access Publisher, 2012

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. 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

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

Ehrig et al. 2003
Ehrig, Marc; Tempich, Christoph; Broekstra, Jeen; Van Harmelen, Frank; Sabou, Marta; Siebes, Ronny; Staab, Steffen; Stuckenschmidt, Heiner: *SWAP: Ontology-based Knowledge Management with Peer-to-Peer Technology.* In: *Wissensmanagement*, 2003, pages 17–20

Ekelhart et al. 2006
Ekelhart, Andreas; Fenz, Stefan; Klemen, Markus D.; Weippl, Edgar R.: *Security ontology: Simulating threats to corporate assets.* Springer, 2006

Gerber et al. 2008
GERBER, Aurona; MERWE, Alta Van d.; BARNARD, Andries: *A functional semantic web architecture*. Springer, 2008

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

Goyal & Fussell
GOYAL, Nitesh; FUSSELL, Susan R.: Effects of Sensemaking Translucence on Distributed Collaborative Analysis.

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

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

Hepp et al. 2009
HEPP, Martin; RADINGER, Andreas; WECHSELBERGER, Andreas; STOLZ, Alex; BINGEL, Daniel; IRMSCHER, Thomas; MATTERN, Mark; OSTHEIM, Tobias: *GoodRelations Tools and Applications*. In: *Poster and Demo Proceedings of the 8th International Semantic Web Conference (ISWC 2009), Washington, DC, USA*, 2009

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

Kingston et al. 2004
KINGSTON, John; SCHAFER, Burkhard; VANDENBERGHE, Wim: Towards a financial fraud ontology: A legal modelling approach. In: *Artificial Intelligence and Law* 12 (2004), Nr. 4, pages 419–446

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

Lara et al. 2007
Lara, Rubén; Cantador, Iván; Castells, Pablo: Semantic web technologies for the financial domain. In: *The Semantic Web.* Springer, 2007, pages 41–74

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

Ozturk 2010
Ozturk, Ozgur: *Introduction to XMPP protocol and developing online collaboration applications using open source software and libraries.* In: *Collaborative Technologies and Systems (CTS), 2010 International Symposium on* IEEE, 2010, pages 21–25

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

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

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

Scharffe et al. 2011
SCHARFFE, François; FERRARA, Alfio; NIKOLOV, Andriy: Data linking for the semantic web. In: *International Journal on Semantic Web and Information Systems* 7 (2011), Nr. 3, pages 46–76

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

Stollberg & Strang 2005
STOLLBERG, Michael; STRANG, Thomas: *Integrating agents, ontologies, and semantic web services for collaboration on the semantic web.* In: *Proc. of the First International Symposium on Agents and the Semantic Web, AAAI Fall Symposium Series Arlington, Virginia*, 2005

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

Verborgh & De Roo 2015
Verborgh, Ruben; De Roo, Jos: Drawing Conclusions from Linked Data on the Web: The EYE Reasoner. In: *IEEE Software* (2015), Nr. 3, pages 23–27

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

Vis.js
Vis.js: *vis.js showcase*. http://visjs.org/showcase/index.html

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

Werner et al. 2014
Werner, Max J.; Vogt, Christian; Schmidt, Thomas C.: *Let our browsers socialize: Building user-centric content communities on webrtc*. In: *2014 IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW)* IEEE, 2014, pages 37–44

Wood et al. 2014
Wood, David; Zaidman, Marsha; Ruth, Luke; Hausenblas, Michael: *Linked Data*. Manning Publications Co., 2014

Yang & Chen 2008
Yang, Stephen J.; Chen, Irene Y.: A social network-based system for supporting interactive collaboration in knowledge sharing over peer-to-peer network. In: *International Journal of Human-Computer Studies* 66 (2008), Nr. 1, pages 36–50

Zhou et al. 2013
Zhou, Yujiao; Nenov, Yavor; Grau, Bernardo C.; Horrocks, Ian: Complete query answering over horn ontologies using a triple store. In: *The Semantic Web–ISWC 2013*. Springer, 2013, pages 720–736

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

Appendix