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, we can assume that e-commerce fraud will still happen in the future and that retailers have to collaborate with relative parties on the incident to find a common ground on and take coordinated (legal) actions against it.

Combining information from different stakeholders will face issues due to different wordings and data formats of the information, 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. Additionally, as some of the information might be confidential or business-critical to one of the involved parties a *centralized* system (e.g. a service in the cloud) could **not** be used.

This Master thesis is therefore looking into the topic of how far a computer supported collaborative work system based on peer-to-peer communication technologies and shared ontologies 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 section of the Master thesis will first give a section showing the importance and relevance of the topic in the research area of Web Science, followed by a description of the problem this thesis will focus on as well as an analysis of related works and an overview of the outline of the thesis.

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."*

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 her 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 card. Additionally, a consumer might hand over her credit card information to an untrustworthy individual, who might use this information for her own benefit. In the real world scenario there is usually a face-to-face interaction between both parties. The consumer, wanting to do business with a merchant or interacting with an employee of a larger business, has to hand over her credit card information explicitly and can deny doing so if she faces a suspicious situation. The criminal on the other hand must get access to the physical credit card first, before she is 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 criminal uses to make copies of credit cards she gets her hands on, or they can be installed in or attached to terminals the consumer interacts with on her 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 of this 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 like card number, card holder and security code for the card validation process – as of this these interactions are usually called *card-not-present transactions*. This credit card information can be obtained by a criminal in a number of ways. First she 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 navigating the consumers to a malicious web site with the intend to capture credit card or other personal information (Consumer Action 2009). Additionally, criminals can **break into the web sites** of large Internet businesses with the goal of getting access to the underlying database of customer information, that in most cases also hold 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/or personal information in this way (Captain 2015).

Based on this it should come not 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 results 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 (mainly 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 online- and mobile 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 inte-

grate 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 such a complex and interconnected system like 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 **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] Instead 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.

Stakeholders might include **vendors** and other businesses, that the retailer has a long-term business relationship with, **law enforcement agencies**, **acquirers** like PayPal or Visa, and even **competitors**, that are also affected by the Internet fraud. In such a case the merchant usually tries to solve the issue on his own and getting in contact with relative 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 involved stakeholders might be a good fit, but arranging such a meeting (same time, 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 the crime the more difficult it will become to find the criminals and take legal actions against them, which can also reduce the risk of losing the stolen money completely.

[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)

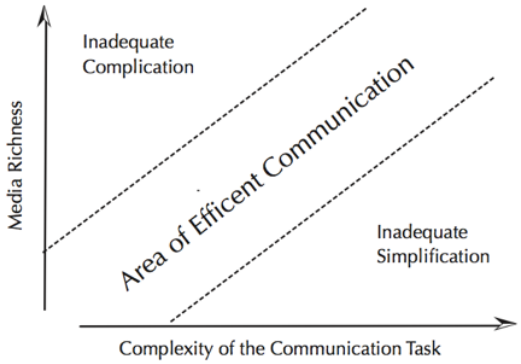

Figure 1.1: The Media Richness Model (Rice 1992)

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

- **communication:** two-way exchange of information between different parties
- **coordination:** management of shared resources like 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 (see Figure 1.2) (Koch 2008):

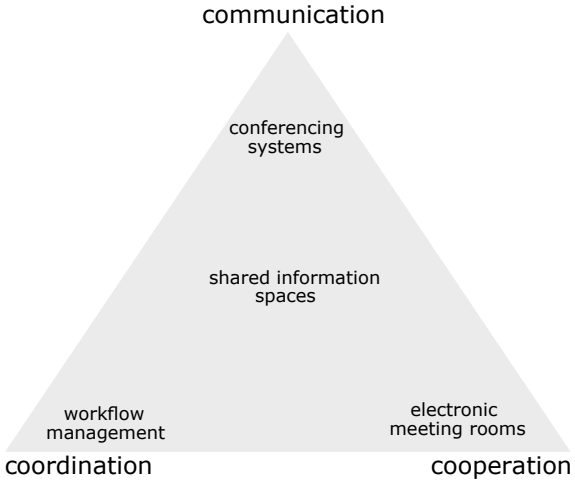

Figure 1.2: The 3C Model (Koch 2008)

A good candidate *could* be a **shared information space**; aka team rooms, cloud storage services or document management systems, that allow to access information at any place, any time and to share information with co-workers — 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) could **not** be used in the scenario described here. Another key characteristic of the investigation of an e-commerce fraud is, 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 the common group goal on time. Combining information from different stakeholders will face issues due to **different wordings and data formats** of the information, **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, that utilizes **peer-to-peer communication technologies**, are either restricted to a commonly agreed set of data entities and relations (based on an ontology) between all involved parties 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, 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:

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

1.3 Master Thesis Outline

2 Context Analysis

ca. 15 pages

This chapter will look into the scenario of e-commerce fraud investigation in detail. It will start with an in-depth scenario description followed by an analysis of all involved stakeholders. It will further describe the kind of information each stakeholder has in her local context and her objectives to take part on the information sharing and collaboration initiative. The chapter ends with a description of the scope this Master thesis will focus on.

2.1 Scenario Description

- different forms of e-commerce exists:
1. Business-To-Business (B2B): electronic trading between companies for improving supply-chain processes
2. Business-To-Consumer (B2C): electronic trading between merchant and consumers
3. Consumer-To-Consumer (C2C): electronic trading between consumers - e.g. eBay

- focus of this thesis is on B2C
- consumer is using an e-commerce shop of a merchant on the Internet
- she is ordering products from the catalog of the merchant and uses her credit card to pay the bill
- the merchant is relying on a third party service to handle the payment process (e.g. PayPal)
- this payment processor is routing the transaction amount due to the issuer of the credit card (e.g. a bank or credit card institute)
- the merchant have a business relationship with her own bank and uses the service to acquire the outstanding amounts from consumers
- in the clearing process the acquiring bank is settling any outstanding transaction with an issuing bank

Workshop ErsteBank Wien:
- usually banks are monitoring the usage of credit / debit cards and are looking for

suspicious activities
- this fraud prevention mechanism is working on a rule-based (non self-learning) or score-based (self-learning) software from a third party
- the outcome of the fraud prevention could be: yes (this is a fraudulent transaction, pls. block it), no (everything looks fine, pls. continue with the process) or maybe (uncertain, pls. let a human decide how to proceed)
- in case of the maybe result an alert is triggered to one of the support staff of the bank (operating 24/7/365)
- still this kind of fraud prevention mgmt. can not solve all issues due to the amount and frequency of the transactions, there is generally a fraud-to-sales ratio of max. 0.11 percent in the EU (meaning 1 promille of transactions are fraudulent)
- still the success rate of fraud prevention is rougly 70-80 percent, means of the fraudulent transactions nearly 80 percent are blocked correctly
- for the rest: the consumer has to actively trigger an investigation, if the case is valid usually the issuing bank will cover the cost (in case of larger amounts an insurance will take over).
- the bank is responsible for pushing the consumer to file a case at police
- most of the filed cases could not be resolved
- there is no court decision yet in which circumstances the consumer might be guilty as well

- ca. 85 percent of frauds are e-commerce frauds (EU: 70-90 percent). Hotspots are Germany, France and US. Frauds are coming from Travel-Shops or Online Merchants and the amount is on average between 500-600 EUR
- e-Commerce frauds will usually not filed at police; in most cases the acquirer is in charge to handle the issue
- if it is known that a merchant has been hacked the bank is usually issuing new credit cards to all affected consumers automatically
- otherwise the bank has to get in contact with the acquirer and/or merchant to figure out if the transaction is fraudulent
- usually the banks only have credit card related information from a consumer (no detailed information about the ongoing transaction), whereas the merchant and the acquirer have the detailed records of the order at hand
- various regulations make it hard to shre detailed information with involved parties (even if they have special agreements signed between them)
- main questions for e-commerce fraud: who is the party that is the victim of the incident? Is it really a fraudulent transaction?
- e-commerce fraud can not be handled by technology alone, at best the fraudulent

transaction can be blocked on the merchant side (due to the information given by the consumer like items, prices, delivery address, ...)
- in the worst case one successful fraudulent transaction in an e-commerce shop will trigger hundred and thousands of attempts -¿ so the awareness for the issue has to be at merchant side
- at the end: much effort is assumed to bring all the experts together and solve the issue by putting their individual know-how on the table

2.2 Stakeholder Analysis

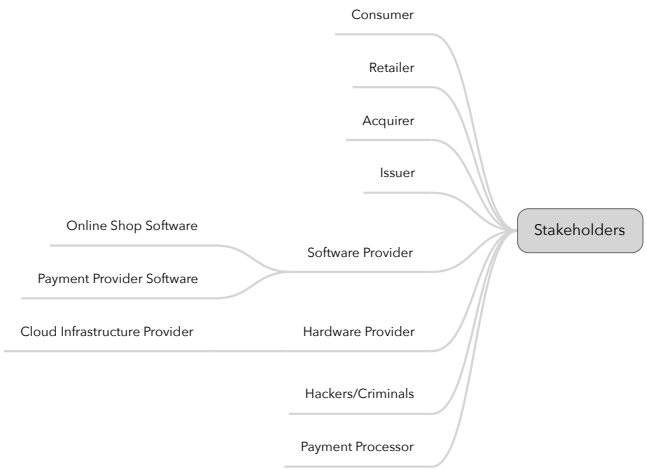

Figure 2.1: Stakeholder Overview

2.3 Stakeholder Objectives

2.4 Scope of this Master Thesis

3 Theoretical Foundations

ca. 25 pages

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, WebRTC as well as less known ones like BitTorrent and BitMessage.

3.1 Computer-Supported Cooperative Work

3.1.1 Definition

3.1.2 Types

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

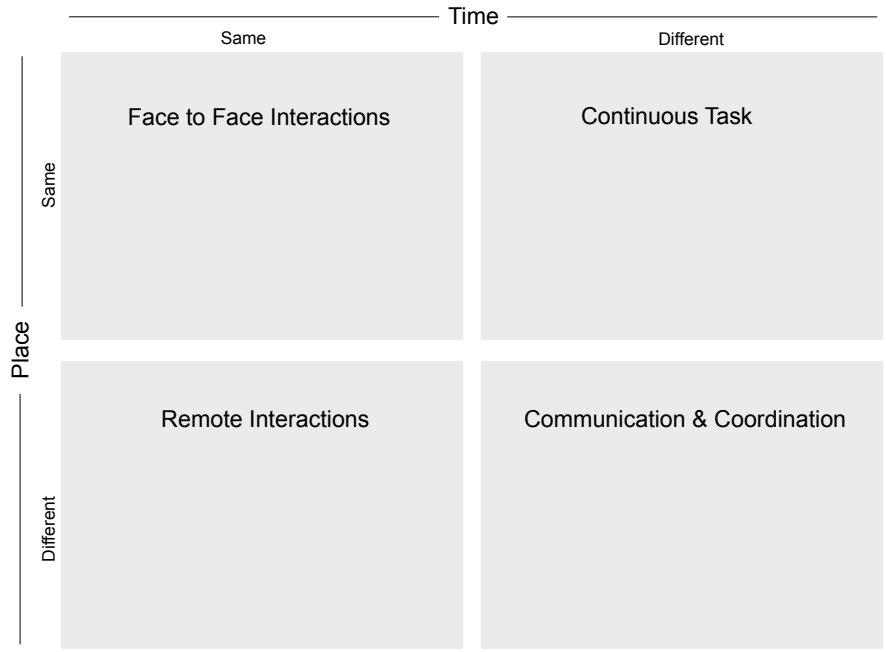

Figure 3.1: CSCW Place/Time Matrix (**?**)

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

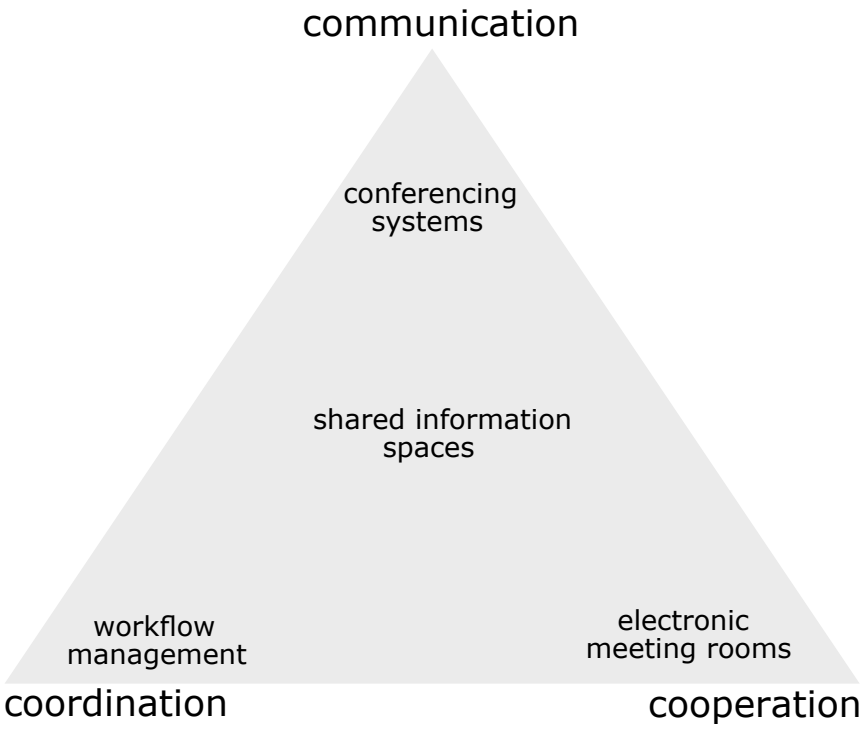

Figure 3.2: The 3C Model (Koch 2008)

3.1.3 Shared Information Spaces

3.1.4 Important aspects of CSCW systems

3.2 The Semantic Web

3.2.1 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 ...

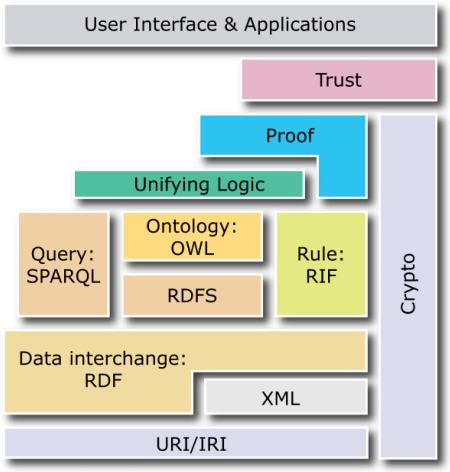

Figure 3.3: The Semantic Web Model (W3C 2013)

3.2.2 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

3.2.3 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> [...]

sample.ttl:

```
@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

sample.xml:

```
<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

sample.html:

```
<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 ()

3.2.4 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)

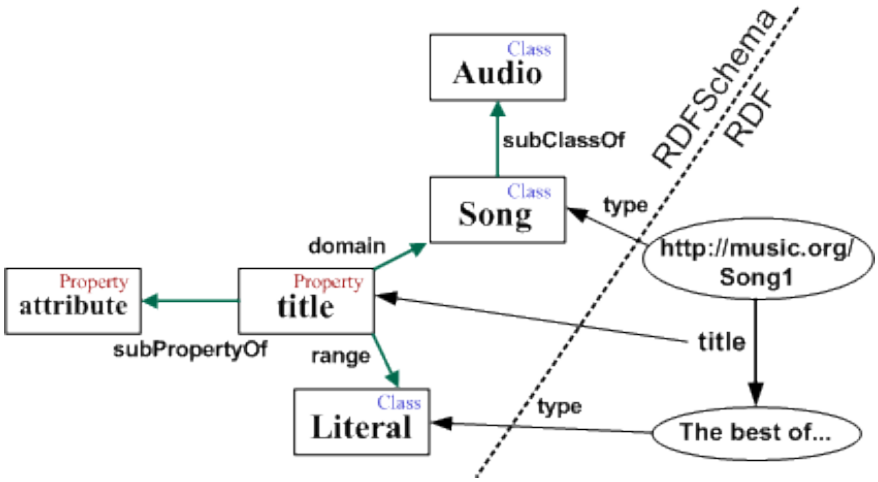

Figure 3.4: RDF Schema sample

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):
...

3.2.5 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:
`sample.sparql:`

```
1  PREFIX ns1:<URI>
2  PREFIX ns2:<URI>
3  PREFIX ns3:<URI>
4
5  SELECT ?varName
6  WHERE {
7      ns1:subject ns2:predicate ?varName
8  }
```

- 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
`sample2.sparql:`

```
1  PREFIX ns1:<URI>
2  PREFIX ns2:<URI>
3  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)

3.2.6 Agents and Rules

3.3 Peer-to-peer communication

3.3.1 Centralized vs. Decentralized Web Architectures

3.3.2 Initiating a communication session

3.3.3 Finding communication peers

3.3.4 Transmitting Data

3.3.5 Available Protocols

4 Related Works

5 Concept and Design of the System

ca. 30 pages

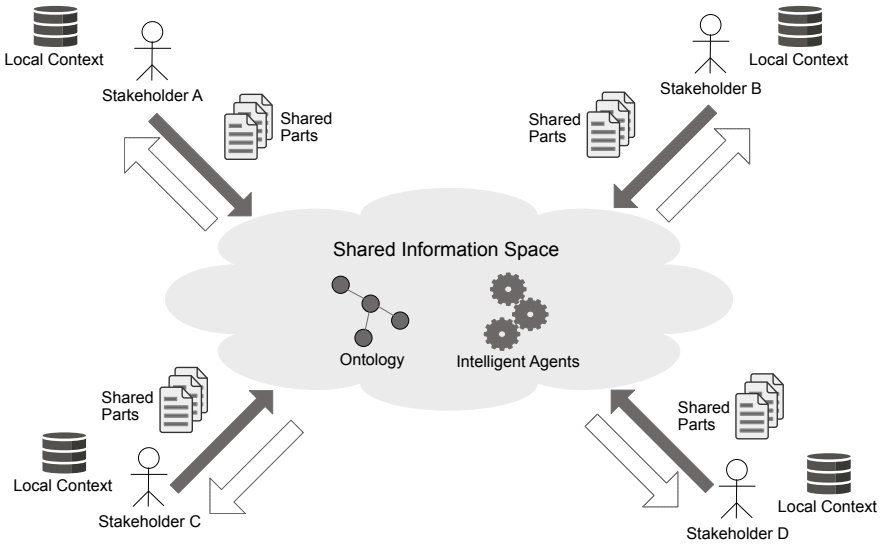

Figure 5.1: System Overview

Based on chapter 3.1 we can conclude:
1. Face-to-Face Meetings: out-of-scope of this thesis
2. Distance Meetings: lack of collaboration support
3. Continuous Tasks: collaboration in teams, but only works when everyone is online
4. Communicate & Collaborate: allows to work on it in a disconnected mode, but increases communication and coordination efforts as well as might lead to synchronisation issues over time

This either leaves us with two options:
1. build a distributed, synchronous collaboration system, in that ppl. can share and work on content at the same time
2. build a distributed, asynchronous collaboration and communication system, in that ppl. can work on things for themselves and get connected together at a certain point in time for synchronising their findings and develop new insights

In the first variant it can be assumed that:
- stakeholders will initiate a collaborative session for a certain case, the collaboration and information sharing efforts end with finishing the case.
- each stakeholder might just work on his part of expertise in the whole knowledge graph (e.g. named subgraphs per stakeholder). these parts could be easily mirrored on the stakeholders environment (no discrepancies with informations from others)
- the whole knowledge graph is only available during the p2p collaboration session, nevertheless results and findings (per stakeholder?) can be synchronized into the named graph of the stakeholder and be analysed offline
- ...

In the second variant it can be assumed that:
- every stakeholder holds different parts of the whole knowledge graph, even might hold the whole graph on his machine.
- stakeholders can fill out the information offline, they might get together at irregular intervals to synchronise their efforts and come up with new knowledge graph entries based on the work of the others
- during the synchronisation process there might come up discrepancies due to the different understandings of the stakeholders for a certain aspect of the knowledge graph
- there might also be different findings or result, even contradictory statements, based on the different progress of each stakeholder on the knowledge graph
- ...

6 Conclusion and Future Work

ca. 10 pages

List of Figures

List of Tables

Glossary

CSCW	computer-supported cooperative work.
OWL	Web Ontology Language.
P2P	Peer-To-Peer.
RDF	Resource Description Framework.
SPARQL	SPARQL Protocol and RDF Query Language.
WebRTC	Web Real-Time Communication.
XMPP	Extensible Messaging and Presence Protocol.

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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