UNIVERSITY OF TARTU

Institute of Computer Science

Software Engineering Curriculum

Ioane Sharvadze

Implementation of WebWeaver Platform over P2P Network

Master’s Thesis (30 ECTS)

Supervisor(s): Vishwajeet Pattanaik, PhD

Tartu 2018

**Title**

**Abstract:**

In the era of Web 2.0 we are using numerous applications that store and manage user data. Switching one application with competitor seems a problem, since we are losing data. While there are numerous approaches to this problem, we propose P2P solution, that gives users real ownership of the data. The goal of this thesis is to show how we enable to have similar functionality of centralized database while storing data on user’s devices.

This thesis will show solution for this particular application WebWeaver, that enables users to add a comment on a web page and share with other users, without website having this feature.

Keywords:

Social blogging, Liquid democracy, Digital divide, Peer-to-peer network

Table of Contents

[1 Introduction 4](#_Toc504859272)

[1.1 Goals 4](#_Toc504859273)

[1.2 Importance 4](#_Toc504859274)

[1.3 Requirements 5](#_Toc504859275)

[1.3.1 Private Share 5](#_Toc504859276)

[1.3.2 Public Share 5](#_Toc504859277)

[1.3.3 Private Shared 6](#_Toc504859278)

[1.3.4 Multiple Devices 6](#_Toc504859279)

[1.3.5 Social Sharing 6](#_Toc504859280)

[1.3.6 Edits, Deletes 6](#_Toc504859281)

[1.3.7 Security, Integrity 6](#_Toc504859282)

[1.3.8 Technical 6](#_Toc504859283)

[1.4 Comparison & Overview 6](#_Toc504859284)

[1.4.1 Musubi 6](#_Toc504859285)

[1.4.2 Linked Data, Solid 7](#_Toc504859286)

[1.5 Limitations 7](#_Toc504859287)

[2 Proposal 8](#_Toc504859288)

[2.1 Why Server? 8](#_Toc504859289)

[3 Implementation 9](#_Toc504859290)

[4 Outcome & Conclusion 10](#_Toc504859291)

[5 References 11](#_Toc504859292)

[Appendix 12](#_Toc504859293)

[I. License 13](#_Toc504859294)

# 1 Introduction

## Goals

This thesis is about enabling Peer to Peer data sharing between users. The goal is to remove centralized server as data store and empower users with real data ownership. Ultimately, we would like to remove server at all and make users communicate with each other without any middleware, but later thesis will show why this is not possible.

This thesis will concentrate building solution for a specific application, called WebWeaver. This is a tool that enables users to weave data into specific web element without knowledge and support of the website. WebWeaver is currently being developed as Chrome Extension, that adds this feature to all websites. This thesis will create simple library, that enables application to store data on user’s device and directly share it with specific users. For establishing communication between users, server is needed, but actual communication and data transfer happens on Peer to Peer network.

## Importance

This section will explain the importance of this research.

There are two major problems associated with centralized data management. Data Ownership and privacy. Whenever using of any well-known applications such as Facebook, Google Plus, Twitter etc. we see that they own and manage our data and it is impossible to continue using other application without losing account data [1]. Imagine after years of using Facebook, user wanted to use another application, that would have similar features, but different interface, or provide extended solution. User would not be able to transfer its data to the new website. Data privacy is also concern for users. While people send messages to each other, their data is stored on third party servers.

While there are different proposals to tackle this problem, yet none of them are widely adopted [1].

While area of decentralized data management remains in active research, topic should be considered as important.

## Requirements

As stated before, this thesis will build this solution for a particular application, called WebWeaver platform. To model adequate solution, we need exact requirements that WebWeaver platform [2] has stated for data management layer. For that we need to understand what application does.

WebWeaver is an annotation tool that enables users to leave a comment on a web comment. They can mark comment as public, private or share it to specific users. WebWeaver uses robust algorithm to detect anchors on web pages [3]. Unlike other web annotators like *hipothes.is* [4] or *genius* [5] WebWeaver aims to improve anchoring strategies in case of dynamic content.

To explain the anchoring problem, let’s describe dynamic page like Facebook. Once I enter the page I will see a different content then the user with different account. In this case if I annotate any element, the other user (or even me on next day) might now see same element at all. This is still an open issue for annotators, because of its complexity. But WebWeaver tries to solve it using similarity calculation [3].

This thesis goal is to provide communication mechanism for WebWeaver, so anchoring algorithms will not be discussed here. But still as for requirement proposed database solution will have to transfer data of a small size, as for anchoring algorithm metadata is relatively small and average comment will require less then 10 kilobytes in size. That means that data sharding and parallel downloads like BitTorrent [6] protocol does, will not be needed, as for parallel downloads matter only when data to be transferred is large enough, because establishing P2P connection itself is slow [7] due to NAT traversal technics.

WebWeaver is a social application, that means that it can share data to other users, make it public or leave it as private. That means security and privacy should be considered.

### 1.3.1 Private Share

Imagine Michael added a private annotation on a web page on his one computer. In this case, only Michael should be able to see this annotation.

### 1.3.2 Public Share

Michael added a public annotation on a web page, that means that everybody in the network should see the annotation when entering same web page.

### 1.3.3 Private Shared

This is the scenario when Michael added annotation and wants to share with his friends. In this case, only friends should be able to see annotation.

### 1.3.4 Multiple Devices

Note that people might have different devices. So, in the first case when Michael added annotation, his other devices with same account should access annotations.

That means that solution in this thesis should handle multi device accounts.

### 1.3.5 Social Sharing

Another Requirement of the application is that, if Michael publicly published annotation and now other users have seen annotation, new user might download annotation from another user.

### 1.3.6 Edits, Deletes

This is one of the hardest case to enable editing and deleting items. After feed is distributed in network and author decides to edit or delete, it should be updated for other users too.

### 1.3.7 Security, Integrity

While network is distributed, it should be possible to validate by another user that annotation that was downloaded is unequivocally created by the specified author. Also, private annotation should not be accessible from other users.

### 1.3.8 Technical

Technical requirements are that, library should work well with Chrome browser and should be able to be integrated in Chrome Extension, as far as WebWeaver is a Chrome Extension.

## Comparison & Overview

This section will overview different approaches for decentralized data management and will show in what sense is our approach better.

### 1.4.1 Musubi

Mobile social application platform Musubi [8] proposes Trusted Group Communication Protocol to send encrypted data with server relay. This is very mobile application based solution that uses server to transfer data, because some of the mobile networks (3G), do not allow incoming connections [8].

Musubi users can send their data only after public key is shared among other users. Its goal is to provide a Framework to develop server less mobile apps. While it’s goals seem to be like ours, it is restricted to group sharing, while in our requirements must support public data sharing, so that people who do not know each other and are not friends can still share some data.

### 1.4.2 Linked Data, Solid

Solid is very interesting platform that was built specifically to target data ownership. It uses Web ID to identify users. User should choose its own data server. Service can be third party or self-hosted, but it should implement Solid interface to support all the features. It uses RDF-based resources to link data [1].

The reason why we don’t want to go with Solid is that it still stores data on non-user device. It can be hard for user to set up service, or find any free server. So, it could be better if without any configuration, user could start using application, but also have own data on personal computer.

## Limitations

Research showed that it is impossible to establish Peer-to-peer communication between users without third party server [7]. The reason is that in real environment, most of devices are hidden behind Network Address translators (i.e. NAT). They give user a temporary IP for communication, they only allow traffic from IP’s that user has requested information.

That means that if user has open IP on its device, NAT will not give any public IP, or even if it gives, it won’t allow incoming connection requests.

To overcome this problem, peers should start requesting connections to each other simultaneously. In that case NATs will most likely (64% times for TCP connections) enable Peer-to-Peer connection. This technique is called Hole Punching [7].

Unfortunately, that means for us that in some cases it is not even possible to send data to peer without server. That case should be always considered and as a fall-back mechanism, relay should be allowed. That means that if Michael wants to send a data to Dwight, both peers will need to connect to the public server (that obviously is not behind the NAT) and first Michael will send data to server and server will redirect the data to Dwight.

# Proposal

In this section, I propose solution, while in next section, implementation will be started based on this proposal.

## Why Server?

Let’s revise limitation. Initial research revealed that, it is impossible to implement server-less application, that will connect users and make it possible to transfer data. There are several reasons:

### P2P connection establishment

To make a peer to peer connection we need a signalling server, because peers must know when another peer wants to connect. Before establishing connection, client data should be sync, ports opened, and metadata transferred. This topic has been long researched, that’s why Google standardized peer to peer connection and created WebRTC (web real time communications) protocol [9]. WebRTC is a set of APIs that is implemented by most of the contemporary browsers. It has been implemented first in Google Chrome browser and thus it fulfils our requirement. WebRTC is just an API, for our solution some signalling implementation should be used. For the sake of implementation, I propose to use *Socket.io* [10]because it is simple, well documented and widely popular library.

### Public data holding

In order to share public data, it needs to be stored somewhere on the server, otherwise for example if user requests annotations for website it has to ask to all users. It’s obvious that asking all users is not a feasible solution.

### Sharing data when peer is Offline

Another reason for using server-side application is that users might not be online when data is shared. Imagine scenario when Alice shares something with Bob, but Bob is not online. In this case when Bob comes online, Alice might not be online. Without server it would not be possible to shared data unless both of them are online. If we have centralized server this scenario would easily be fixed. Alice would leave a message to Bob, when Bob gets online he would check centralized server for messages addressed to bob and download message.

## Server Architecture

### Goals

Before starting discussing server, I want to build it with several goals in mind:

1. Server is not application specific.
2. Server’s APIs can be implemented differently.
3. Server should not be trusted.

These goals arise from requirements. As we want to give users freedom and data ownership, server should have least knowledge of application specifics. To support that ultimately, we can hide that information at all. That means that server just hosts any kind of application and does not have knowledge about hosted data and it’s usage.

Because applications are not coupled on server implementation, it will be possible to use different servers with different implementation that support same APIs. That will give server maintainers freedom to choose how much data they allow to hold, whom they allow to communicate or if they add additional security checks.

By nature of this solution, if server does not know about data and cannot check data integrity, users are in charge of validating sender of the data and integrity.

To sum up server is very generic, that can serve any type of application, without any registration needed. Server is not trusted and it should not know anything about the data that is shared. Also, application should be ready to work in case server fails to verify sender and data integrity.

Similar idea of application was mentioned in Musubi [8] that enables developers to use common servers to share data between users. It also made sure that data is encrypted, so that users could verify sender and data integrity. But unlike our solution it was designed for mobile phones and lacked peer to peer data sharing support.

### Idea

To accomplish requirements of Peer to Peer data sharing, we need to create a service that can help peers to establish connection, that service is called “Live Rooms”. Second Service that helps to communicate with offline users is called “Message Box”. First service is for establishing peer to peer connections, second is just temporary store of information, like a real-life post office. Here are scenarios in order to better understand proposed architecture.

If Alice wants to send data to Bob, first she should check service called “Live-Rooms” if Bob is online. If Bob is online, Alice and Bob start sending signalling metadata in order to establish peer to peer connection. If Bob is not online, then Alice can leave a message in centralized storage i.e. “Message Box”.

Bob will synchronize “Message Box” once he comes online.

These two services are not connected together, thus application can choose different service providers, or support only one of them.

Our services use unique key in order to distinguish between users.

## Message Box

### Functionality & Technology

The first step of implementing server is a message box. As explained above it’s a simple database where it holds messages.

Server will have three required functionalities.

1. List Message Id’s for User.
2. Download Message by message Id.
3. Save Message and assign unique Id.

These three functionalities are all that is required. Some servers might extend functionality and add authentication, that will force users to prove that they are indeed who they claim to be.

To model server, I used most commonly used client-server architecture - REST [11]. Because it is most commonly used and accepted approach to model server that holds a data and wants to expose to wide public. Also, technologically it is possible to implement client for browser, mobile and other servers. You can see API in Appendix

### Extended functionalities

As said before, servers are free to extend functionality and some of the servers might choose to different strategies in order to evict the data.

Some might enable only chosen public keys to leave information. This might be chosen for companies that want to enable only small group of people to use the server.

Public servers will most likely implement time invalidation mechanism, where message is deleted after some time, or when there is not enough memory to save new messages. This would make sense, because this service is meant to hold data temporarily, to just synchronize messages.

Some servers might set limits for users who have too many notifications can particular user leave.

Basically, anything should be possible unless they provide 3 basic functionalities.

## Live Room

As Explained before, Live Room is a place where data can be sent P2P. Live Room would handle all cases if all the users were online all the time. But because in real live people are not always on, we need to extend service with Message Box.

The idea of Live Room is to enable fast (real-time) data sharing between users with P2P traffic. Some applications might prefer to only user Live Room, because no server will ever hold the data. But still data encryption is absolutely required, as for by nature of P2P connection establishment [7], data can be easily leaked.

As for Live Room, it will be place where online users gather and wait for incoming connections. Each user will connect to its own “Room” that will be associated with user id. Whenever anybody wants to connect with user, they will connect in this room and ask for connection.

After asking, user will decide whereas they want to connect or not. Some applications might choose to block connections from unknown users, while some might allow. Connection blocking will be decision for application.

After users are connected, they can send data, verify delivery and close connections.

## Client

In order to integrate functionality in application, library client should be written. The library should be compiled in single JavaScript file that can be added in html file as a script file. In this way, library user can use functionality with simple API.

Client Library should not expose inner implementation to the user, but provide an abstraction to save, query public and local data.

### Configurations

Library user should be configurable in order to satisfy different kinds of application needs. Above, I already mentioned that there can be different kinds of servers that extend basic functionality and provide more security and robustness, that is why we need a way to change configuration from the client library.

For the initial version of client library following parameters can be configured:

* “Live Rooms” URL
* “Message Box” URL
* “Message Box” synchronisation interval
* Local Database Name
* “Live Rooms” WebRTC configuration that contains: STUN [9] server URLs.

All the parameters can be left as optional, because everything can be left as default.

In case of “Live Rooms” and “Message Box” URLs, default public service URLs will be used. The services will be maintained by me on Heroku Cloud Application platform [12], thus they will have strict limitations on memory and performance.

Local Database name will be generated with default name, that will be generated based on user Id. This feature is important since, database should be changed upon user Id change. Imagine the scenario when user logs in with different user Id, in this case different database should be used, to avoid reading other users data. In case of automatic database name, application will use corresponding user’s database.

“Live Rooms” WebRTC configuration is needed to change default Stun server URLs, that help to establish peer to peer connection. This might be needed to be changed in case library user wants to use its own STUN server to improve performance and security. Otherwise application will use default STUN server maintained by Google.

### API/Usage

This section describes Client Library API and explains it’s behaviour.

In order to start using library in web application, user has to import library script file using JavaScript Script tag.

User will be able to use those methods:

* sync()
* getPublicDataByKey()
* getByKey(key, callback)
* getByAuthor(key, callback)
* saveData(data, sharedWith, callback)
* listenDataChanges(callback)

Sync method provides a way to force synchronisation between “Message Box” and application. This is useful if application needs to get fresh data and cannot wait for scheduled update interval. In Applications, this might be a case when user forces synchronisation by clicking button.

Get public data by key is useful when library user wants to read public data. The library will try to fetch the “Message Box” for any available public data, but also will check local storage in order to provide the data.

Get by key will be used to retrieve data with application defined key that is a String Id. Note that Key is not required to be unique, so user will receive list of results in the call back, or null if nothing was found.

Get By author works the same way as get by key, but the results are queried by Author Id, that is a public key of the message author.

Listen Data Changes is a convenience method in order to keep application state in sync with storage. If somebody sends a message via “Live Rooms” or “Message Box” retrieves new messages from the cloud, users will get new data objects in call back. This should be fine place to query database again and display new results if needed.

### Background Sync

Library should be designed to work both with simple Web applications and Chrome Applications. The difference is that Chrome Applications can store unlimited data and have the ability to synchronize data even if user has not opened application.

Chrome extensions have a notion of Background Pages, [13] that empower application developers to run scripts while page is not visible or open by users.

The library should use Background Pages in order to get “Live Rooms” connections and share/receive data, but also to fetch new messages from the “Message Box” when scheduled time arrives.

In case of regular web applications, no background synchronisation will happen, but only while application is opened in browser tab.

# Implementations

This section will describe actual technical implementation of each project.

“Live Rooms”, “Message Box” and “Client” projects will be discussed. As described above, they are distinct projects, so I created 3 separate projects that don’t share any code. This will make sure to reduce coupling and enable extending functionality without modifying other projects.

## Live Rooms

Live Rooms is built with “Socket.io” library. It was chosen because it is widely used and designed to work with web applications, therefore it has 2 separate parts. Web Client and Server Client. Here we discuss “Live Rooms” server part only.

“Live Rooms” is written with JavaScript Node.js framework. While other implementations can be done with different JavaScript web frameworks, this is the easiest and most popular choice among the developers [14].

I chose socket library instead of other popular choices like REST for this particular problem, because there are bidirectional events that can be transferred, between server and client. We have a scenario where client is waiting for a server event, that might happen while client is connected. Particular case is when user is waiting for peers to connect.

For this reason, client needs to have a connection with server, so that server can notify client when other peer requests connection. Other solutions like pooling could be applied in this scenario, but the problem is that clients will have to send several consequent signalling messages through the server, because of that managing several signalling messages with pooling approach [15] would be less efficient and harder to maintain. Socket implementation directly solves the problem where many bidirectional messages need to be transferred.

“Socket.io” library has a notion of events that occur. The first event that happens when user connects the server is event named *connect*. It means that user has a ready socket that can be used to send a message or receive a message from.

### Message Contract

After User is connected with server, we listen to several types of events from user. Every message sent by user has to be a JavaScript object having those attributes:

* fromPublicKey
* toPublicKey
* data

fromPublicKey means a public key of a message sender. This is required to be present in all messages to correctly identify sender.

toPublicKey is only required when message needs to be redirected to the other user.

Finally, any other attributes can be added, for example data. Note that server does not care about other attributes, so as far as client can read the data it can be encrypted for security.

### Events

After user is connected, we listen to socket events.

First event to expect after connection is called “*enter\_my\_room”*. This event fires when user is ready to give information about itself i.e. it’s public key. Server joins user to the room named with user public key. “Room” is a “Socket.io” notation that enables us to label sockets. We need to label sockets in order to check if peer is online and redirect messages from other peers. In this simple way we can have multiple parallel signalling between peers. Imagine if Alice wants to contact Bob and Charlie first she will join room Alice and signal rooms Bob and Charlie. If users are waiting in their rooms, server will redirect Alice signalling messages to those users.

If any event fails, server usually send an event “*error*” to notify client about failed action. Server also sends a failed message, so that client can reset state and try again if needed.

Once user is entered its room, server is ready to notify peer connection requests. User also can initiate peer connections at this stage.

In order to initiate connection request, event “*connection\_request*” should be sent. Note that this event also requires peer public key, so that peer can get an event called respectively “*connection\_request*”. At this stage peer can either accept or ignore request.

By accepting request Client will gather

# Sample Project

Sample project was done in order to show the usability of the Library.

# Security

# Outcome & Conclusion

[9]

# References

|  |  |
| --- | --- |
| [1] | E. Mansour, A. V. Sambra, S. Hawke, T. Berners-Lee, M. Zereba, S. Capadisli, A. Ghanem ja A. Aboulnaga, „A Demonstration of the Solid Platform for Social Web Applications“. |
| [2] | D. Draheim, M. Felderer ja V. Pekar, „Weaving Social Software Features Into Enterprise Resource Planning Systems“. |
| [3] | „Making Web Annotations Dynamically Robust and Semantically Rich“. |
| [4] | „hypothes.is,“ [Võrgumaterjal]. Available: https://web.hypothes.is/. |
| [5] | „genius.com,“ [Võrgumaterjal]. Available: https://genius.com/web-annotator. |
| [6] | J. Pouwelse, P. Garbacki, D. Epema ja H. Sips, „The Bittorrent P2P File-Sharing System: Measurements and Analysis“. |
| [7] | B. Ford, P. Srisuresh ja D. Kegel, „Peer-­to­-Peer Communication Across Network Address Translators“. |
| [8] | B. Dodson, I. Vo, T. J. Purtell, A. Cannon ja M. S. Lam, „Musubi: Disintermediated Interactive Social Feeds for Mobile Devices“. |
| [9] | C. Holmberg, S. Hakansson ja G. Eriksson, „Web Real-Time Communication Use Cases and Requirements,“ March 2015. [Võrgumaterjal]. Available: https://tools.ietf.org/html/rfc7478. |
| [10] | M. Moore, „The semantic web: an introduction for information professionals,“ Thomson Reuters, 2011. |

Appendix

1. License

**Non-exclusive licence to reproduce thesis and make thesis public**

I, **Ioane Sharvadze**,

(*author’s name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

of my thesis

**Implementation of WebWeaver Platform over P2P Network**,

*(title of thesis)*

supervised by      ,

*(supervisor’s name)*

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **10.04.2018**