SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES

Ing. Radovan Semančík

Revised World Wide Web Architecture

Summary
(unofficial English version)

submitted in partial satisfaction of the requirements for the degree of

philosophiae doctor

in

25-21-9 Computer Tools and Systems

Bratislava, December 2008

Submitted by: Ing. Radovan Semančík
Ústav počítačových systémov a sietí
FIIT Slovenská technická univerzita v Bratislave
Ilkovičova 3, 842 16 Bratislava

Supervisor: doc. Ing. Margaréta Kotočová, PhD.
Ústav počítačových systémov a sietí
FIIT Slovenská technická univerzita v Bratislave
Ilkovičova 3, 842 16 Bratislava

Opponents: prof. Ing. Pavol Horváth, PhD.
Centrum výpočtovej techniky, STU, Bratislava

prof. Ing. Jiří Šafařík, PhD.
Fakulta aplikovaných věd, ZČU, Plzeň

Ing. Martin Chovanec, PhD.
Ústav výpočtovej techniky, TU, Košice

Committee chair: prof. Ing. Liberios Vokorokos, PhD.
Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach
Letná 9, 042 00 Košice

Note: This is non-official English version of the original document in Slovak language.

TABLE OF CONTENTS

Introduction

The interaction with computer system is an important part of our lives. The computers frequently store and use data that describe characteristics of physical human beings. Yet only a few attempts had been made to provide a model that would help to understand implications of personal data processing.

This paper provides a basic structure of a model that may provide insight into some of the areas that touch personal data. Some concepts in the problem area are understood only intuitively and this work attempts to provide more formalized definitions and it attempts to put these concepts in broader perspective. The work was motivated by a desire to design an "identity layer" on top of current World Wide Web architecture. However, that approach was changed in the process of modeling the system and analyzing the basic structure of the World Wide Web. We have discovered that basic identity-related mechanism must be integral part of the World Wide Web architecture to be effective. This work describes the motivation and the reasons for such conclusion and it provides a proposal of the necessary changes to the architecture of the World Wide Web.

Today's World Wide Web was built on assumptions that most of the shared information is public. It also expects that only a very limited number of subjects will publish and modify the information while many subjects will read it. However, such situation will not hold for much longer. Many applications build on top of the World Wide Web are actually making it *writable*. Wiki, blogging, photo and video sharing applications allow to add comments to the resource, therefore influencing its state. The original World Wide Web architecture has not anticipated that most of the resources will be dynamic and many of them non-public or semi-public.

1 State of the Art

Password authentication is a traditional method of authentication, however it is inadequate for highly distributed systems [1]. To overcome the security issues of passwords, several One Time Password (OTP) [2] [3] and challenge-response authentication schemes [4] [5] were proposed. However both one time password schemes and challenge-response schemes in general have some common security drawbacks. When used in a plain TCP/IP environment, connection data can be manipulated after a successful authentication takes place. In this case the attacker does not need to attack the authentication scheme directly. These attacks can be prevented only by pre-authenticating the server to the client by using other independent methods and explicitly authenticating the transported data.

The Public Key Infrastructure (PKI) is a set of methods and formats for the management of public keys and all related data. PKI uses asymmetric cryptography methods to achieve its goals. The international standard for the public key certificate format is based on ITU-T X.509 recommendation [6] and is widely used in both enterprise and Internet environments. The certificate authority affirms certificate validity by its signature. Public key certificates can be used in many communication systems on the Internet. The most common communication protocol in use today that employs X.509 public key certificates is the Transport Layer Security (TLS) Protocol [7]. The most frequently used TLS mode today is the authenticated server mode. The server has an X.509 certificate for its fully-qualified domain name (FQDN) and the client (web browser) has a list of trusted certificate authorities. If TLS is used in any authentication mode, it provides only short-term security for the transported data - TLS

protects and authenticates data on the communication channel only. If the data is stored on the target system, they are no longer protected. Even if the data block was received by the server in a mutually authenticated TLS connection, server is not able to provide any proof of data origin to the third party. The most common practical use of the TLS protocol is to secure a WWW communication. It is called HTTPS [8] and it is essentially HTTP protocol communicating over a TLS-secured channel. Most of the connection-oriented protocols can be modified in a similar way to use the TLS layer for protection.

Public key infrastructure is not well-suited to express dynamic trust relationship. The digital certificates used by public key infrastructures are relatively long-lived, they are difficult to update and revoke. Therefore the public key infrastructure is used to express relations that are quite static, such as binding name to an entity or to express contractual relations between organizations. Attempts to use client-side certificates haven't gained any considerable user adoption. A low penetration of qualified certificates as authentication method for government services in European countries is a clear indication of this phenomenon [9]. High cost of the certificates and low benefits may be one of the reasons of low PKI adoption. Except for low adoption there are privacy concerns of PKI usage [10]. Once the certificate is presented, all the attributes and data in the certificate are disclosed to the verifying party. That means that the certificates should contain only minimal amount of information to conserver user's privacy. However, such minimal certificates have limited usability and even such minimal information may expose user to a collusion attack.

Digital credentials [11] is a cryptographic system based on secret-key certificates that was designed to overcome some PKI problems. Although the basic principles of the system are documented, the communication protocols and the details of usage are not. Therefore it is not possible to throughly evaluate digital credentials at the time of writing this document.

Internet identity systems [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] attempt to move the maintenance of user accounts to a dedicated and shared sites, therefore lowering the number of accounts a user has to maintain. The mechanism used by such systems is a secure transfer of user's identifiers, attributes and current authentication status of a user from source site (Identity Provider) to the destination site (Service Provider). The trust relationship must be established between source and target sites for a source site to trust the target site's requests and for a target site to trust the source site's identity statements. Establishing and maintaining this trust relationship is out-of-band for most Internet identity systems.

The Internet identity systems supports one or both of following methods. Browser-based mechanism assumes nothing more than a plain web browser on a client-side takes advantage of HTTP redirects to transfer security tokens between sites. Client-based mechanism assumes existence of special-purpose software component on client-side (identity client) that handles the transfer of security tokens between sites. Internet identity systems follow the proxy-based true SSO model according to Pashalidis and Mitchell [22].

Internet identity system suffers from a variety of problems. Source site can trivially impersonate any user that on target site [22]. The source site is able to track the user's log-ons on service providers sites. While the source site cannot track the user's activity at these sites, the log-ons itself may provide sufficient information to potentially violate user's privacy. Global identifiers used by some systems can be used to correlate user activity as several target sites, endangering user's privacy. Some of the systems (especially OpenID [18]) is susceptible to so-called *phishing* attacks [23], made possible by the architecture, user interface and design

decisions [24]. Internet identity systems either assume direct trust between sites (e.g. expressed by exchanging X.509 certificates) or they assume trust implied from other system. The implied trust is usually sourced from DNS or the public key infrastructure of HTTPS, such as in case of OpenID [18]. However, such approach is not correct as both DNS and HTTPS structures express only name assignment and does not express any kind of trust.

Several authors [25] [26] discuss the conceptual foundation of anonymity and identity in computer systems. The work of Pfitzmann and Kōhntopp [27] is especially noteworthy, building common terminology on the works of other authors. However the work frequently operates with boolean values where a scale would be more appropriate and it does not distinguish between human actors and computer actors in the system. The modern concept of privacy was established by Warren and Brandeis [28] as "the right to be let alone". Solove [29] describes that traditionally view on privacy as an invasion paradigm and he argues that it is no longer efficient to address the privacy risks of 21st century. He proposes architecture based on Fair Information Practices [30] which was a one of the sources for OECD guidelines for the protection of privacy, that are similar to the EU personal data protection directive [31]. Resnick et al. [32] are discussing ad-hoc interactions on the Internet between parties that have no previous relationship. They propose a distributed reputation mechanisms a solution for problems inherent in such interactions. The positive effects of reputation systems on ad-hoc interactions are also described by Axelrod [33], proposing evaluation of past actions for the purpose to be used in future decisions. He describes it as "shadow of the future" that can motivate users to a better behavior in the present by threatening to punish bad behavior in the future. However, Windley et al. [34] note that there is a natural trade-off between reputation and privacy. They argue that as reputation is calculated from the record of past interaction, revealing such record means that part of the subject's privacy is lost. Sabater and Sierra [35] provide a review of computational trust and reputation systems, classifying them by a variety of criteria. They note that the reputation can be seen as global or subjective. A game-theoretic approach to model Internet interactions based on examination of the record of past interactions is provided by Friedman and Resnick [36]. It is shown that the availability of cheap pseudonyms is harmful to the level of overall cooperation in the network. Friedman and Resnick propose the creation of once-in-lifetime certificates to mitigate the effects of cheap pseudonyms.

World Wide Web (WWW) is a distributed global hypermedia system that originated as simple hypertext system in early 1990s based only on a short proposal [37]. The principles and protocols of current World Wide Web architecture have evolved during late 1990s, resulting in definition of URI [38] and HTTP 1.1 [39]. The architecture was guided by the Representational State Transfer (REST) architectural style described by Fielding [40] in 2000. However, there was no architectural document for World Wide Web until 2004, when the *Architecture of the World Wide Web, Volume One* [41] was published by World Wide Web Consortium (W3C). The WWW architecture document retrospectively documented the thinking behind the development of World Wide Web. However according to Fielding [40] the architecture and protocols of World Wide Web we only roughly aligned with the REST architectural style and inconsistencies still remain.

Current WWW browsers does not indicate the trustworthiness of displayed information. The information that browser typically display is limited to URL of the page and indication is HTTPS usage. This leads to a success of *phishing* attacks [42] because users cannot distinguish authoritative authentication page from a fake page with the same design. The usability study

accomplished by Dhamija, Tygar and Hearst [42] demonstrated that good phishing site deceived 90% of participants.

Current architecture of World Wide Web assumes that information always comes from its authoritative source or a trusted proxy. The HTTPS mechanism is designed to be effective for protection of information under such assumption. However, the usability of HTTPS is limited when such assumption does not hold, for example in case of on-demand data replication and migration.

Internet is a world-wide communication medium. It connects almost 1.5 billion of users [43]. It is unrealistic to assume that any significant part of such a huge population will maintain a long-term relationships with all the people they interact. It must be expected that significant part of the interactions will be carried out between people and organizations that does not know each other. Therefore a method is needed that helps Internet users do evaluate the trustworthiness of interaction partners and generally any information available on the Internet. The core of the problem lies in the architecture of World Wide Web. World Wide Web mechanisms does not provide any way that can assist users with evaluating trustworthiness of communication party or reliability of information discovered on the Internet.

2 Dissertation Objectives

The work is focused on the improvement of World Wide Web architecture, especially on conceptual layers of the architecture. The objectives of this work are defined as follows:

- Define the goals and expectations for the World Wide Web architecture and design for a current needs and for a foreseeable future.
- Evaluate the state, consistency and appropriateness of World Wide Web architecture according to specified goals.
- Identify and discuss fundamental problems of WWW architecture, especially focused on handling of user identities and evaluating trustworthiness of commutation parties and reliability of information in WWW environment.
- Propose architectural improvements in World Wide Web architecture on a conceptual level, leading to improved support for evaluation of information reliability.
- Validate the proposed architecture and demonstrate that the proposal can handle situations that are problematic in current World Wide Web.

One of the primary implicit goals of this dissertation was to keep the efforts of designing the improvements of World Wide Web architecture feasible. We sought to provide complete results at the conceptual level that can be reviewed and evaluated rather than incomplete and therefore non-conceptional solution of a single World Wide Web aspect. The review and improvement of World Wide Web architecture is definitely not an easy task, therefore we opted to split it to several steps. This document describe the results of the first step, improving the mistakes of World Wide Web architecture on the conceptual layer. Therefore the objectives of this dissertation are set specifically for that purpose.

3 The Model

No practical architecture can be conceived without first understanding the concepts that underlay the architecture and the needs of the people that seek to gather the benefits of the architecture. Therefore we provide a model that attempts to describe and explain the undercurrents of realspace-cyberspace interactions. The model created for the purpose of this dissertation is based on the interaction of two worlds: the world of human beings (called *realspace*) and the world of computers (called *cyberspace*). We discuss how people deal with computers and the implications on the reliability of information provided by computers. We also deal with anonymity and identity of people in regard to data processing.

The interaction between realspace and cyberspace is made possible by *terminal devices*. These devices are entities that are part of both spaces and they convert information from a form perceivable in one space to the form suitable for the other space. Computer monitor, keyboard, camera or independent sensor are examples of terminal devices. Neither realspace nor cyberspace entities are sure whether the terminal device operates as expected, as they cannot perceive the other world directly. Correlation of information from several terminal devices may increase the confidence in the information, however the reliability of the information always disputable.

Based on the discussion above we can formulate following statement:

Crossing the boundary of realspace and cyberspace is always subjective.

The entity receiving data from other space using a terminal device must make its own assumptions about the relevance of the data. It has to (implicitly or explicitly) evaluate a level of belief that the data describe what they are supposed to describe. We consider realspace-cyberspace interactions to be *subjective* (as opposed to being objective). The interaction depends on the interpretation of the information by both realspace and cyberspace entities, on their preconceptions, predetermined behavior and believes, on the presentation and detection capabilities of terminal devices, on the environment and overall situation of the interaction. As any information that resides in the cyberspace originated in the realspace and had to pass realspace-cyberspace boundary, we may formulate following statement:

Any information coming from the cyberspace is subjective.

Therefore the trustworthiness of the information cannot be reliably evaluated unless its source is known. The credibility of the information source must always be considered to determine the likeness that the information is true. Therefore we can formulate following statement:

The source of the information in the cyberspace is equally important as the information content.

Based on the reasoning above, we do not require the user of information should have complete knowledge about the realspace identity of the source of information. We rather propose that appropriate information about the source should always be conveyed with the information content. We also propose that the source is always taken into consideration when the information is used, while the actual mechanisms of consideration may vary.

Persons (realspace entities) acting as users of computer systems are represented in the cyberspace by data structures. As explained in the previous sections, the computer systems that are interacting with realspace persons have limited capabilities of determining the person

characteristics directly. The data structures that represent users are assembled from subjective information which is commonly entered to the system by the users themselves. The data structure maintained in the computer system is in most cases incomplete representation of realspace person, with variable reliability.

We will use the term *Subject* to represent the the realspace person and the term *Persona* to represent the cyberspace data structure maintained in the computer system that is related to the person:

Subject is a conscious realspace entity that is guided by a free will.

Persona is a cyberspace data structure that represents some aspects of (realspace) subject or a cyberspace entity that is governed by realspace subject. The aspects are represented as a collection of properties with machine-readable values.

As personas a cyberspace data structures, it follows that the cyberspace data structures of persona can be only a subjective representation of realspace subject.

The persona usually originates as an data structure describing one realspace person (subject). But after the original creation, the link between the persona (cyberspace entity) and subject (realspace entity) may not be apparent. According to this, a mechanism is needed that will allow evaluation of relations between personas and subjects in a running system, long after the creation of personas. We define a concept of a world set that is a set of all candidate subjects that could be used as a source for personas in the system we consider, denoted as follows.

$$W = \{S_1, S_2, S_3, \ldots, S_n\}$$

We define basic concept of identity using probabilistic mechanism:

Identity probability is a (Bayesian) probability that given persona describes given subject. Denoted

$$i_{P,S}$$

whereas P is a persona and S is a subject.

The identity probability value, as well as all other probabilistic metrics defined in the document, are subjective to a specific observer and depend on his knowledge. While persona describes exactly one subject (by definition), the sum of identity probabilities for specific persona and all subjects in the world set must be 1. The following holds:

$$\sum_{k=1}^{n} i_{P,S_k} = 1$$

We can define a random variable I_P with the world set W as the collection of source states and with individual probabilities being equal to identity probabilities of a single persona:

$$P(I_P = S) = i_{P,S}, \quad \forall S \in W$$

The random variable I_P represents possible subjects that the persona P may describe.

Given a specific persona, **anonymity set** (denoted AS_P) is a set of all possible subjects from the world set for which holds that the identity probability of the persona and the subject is greater than zero. Can be denoted as:

$$AS_P = \{S : \ S \in W \ \wedge \ i_{P,S} > 0\}$$

This definition of anonymity set is a probabilistic extension of the definition used by Pfitzmann and Kōhntopp [26].

Anonymity ratio of a persona with respect to world set and observer describes the relative uncertainty in persona correspondence to a subject. Defined as

$$ar(P) = \frac{H(I_P)}{H_{max}} \quad ,$$

whereas $H(I_P)$ is an entropy [44] of random variable I_P, defined as

$$H(I_P) = -\sum_{k=1}^{n} i_{P,S_k} \log_2(i_{P,S_k}) \quad ,$$

and H_{max} is a maximum entropy for a random variable with n states, that is

$$H_{max} = \log_2(n) \quad .$$

Anonymity ratio of 1 means total anonymity: the persona may describe any subject from the world set with equal probability. The anonymity ratio will be 1 if the observer cannot in any way distinguish the subject that the persona describes (the probability distribution of random variable I_P is uniform). Anonymity ratio of 0 means no anonymity: The persona describes a single specific subject. The anonymity ratio will be 0 if the observer is sure that the persona describes a specific subject.

Identity ratio of a persona with respect to world set is a probabilistic inverse of the anonymity ratio. Defined as

$$ir(P) = 1 - ar(P) \quad .$$

The identity ratio describes the degree of observer's belief that persona P describes some specific subject. Identity ratio of 0 means no identity: the observer cannot infer any useful information about the subject that persona describes. Identity ration 1 means total identity: the observer is sure about the subject that the persona describes.

Exact values of anonymity ratio and identity ratio may be very difficult (if even possible at all) to compute in the practice. Only the estimated values of these metrics can usually be determined. However we may use other metrics that can be computed in computer environment and may provide estimations of anonymity and identity: analogy and heterology relations and analogy probability value that may be used to approximate anonymity and identity in practical scenarios.

The personas are **analogous** if and only if the observer believes that they describe the same subject.

Analogous personas describe the same subject. These personas may be two accounts in different systems that belong to the same user or two database records describing the same person. The analogy defined in this deterministic manner may not be very useful in practice. It is usually difficult (if possible at all) to reliably decide if two personas are analogous, but it is usually feasible to provide estimation on how likely it is that the personas describe the same subject. For this reason we define the probabilistic version of analogy:

Analogy probability is a (Bayesian) probability that two personas are analogous. Denoted

$$l_{P_1, P_2} \quad ,$$

whereas P_1 and P_2 are personas.

Analogy probability can be seen as a degree of observer's belief that two personas describe the same subject.

The personas are **heterologous** if and only if they describe different subjects.

Similarly to the concept of analogy we define a probabilistic version of heterology:

Heterology probability is a (Bayesian) probability that two personas are heterologous. Denoted

$$h_{P_1, P_2} \quad ,$$

whereas P_1 and P_2 are personas.

Two personas can only describe the same subject or two different subjects. Therefore it follows that

$$l_{P_1, P_2} + h_{P_1, P_2} = 1 \quad .$$

Analogy probability and a world set based on persona databases may provide an estimation of anonymity/identity ratio values in practice.

4 Design Goals and Methods

Architectural model similar to the model proposed by Fielding [40] is used as a guiding principle in this work. The behavior of the designed system is described in forms or constraints: what the system must do and what it must not. Architectural style is formed as a named set of architectural constraints.

Our goal is to propose an architecture that will support an environment of effective cooperation. Such environment should induce the positive network effect [45]. It should encourage the cooperation of any two entities in the network. The cooperation should not be limited to channel-oriented interactions, where few strong entities mediate most of the interactions on the network. It is expected that the environment will change as the society changes. The designed system must address such a dynamic nature of the environment. No information should be regarded as permanent. The dynamics of the information must be taken into consideration and be reflected in the architecture.

Three environment settings are considered: anarchy, authoritarianism and environment of responsibility. Anarchical environments inhibit cooperation. The effect of anarchical environment will be a wilderness where few strong individuals will abuse the majority of others, making them even weaker and more susceptible to domination. The business is

ineffective, as nobody can be trusted. Long-term relationships are necessary to build-up relationships of trust and no long-term relationships could be established if the identity of participants is actively hidden. The authoritarian environments do not scale, therefore are not usable for internet environment. Even if the initial authoritarian regime cared about the well-being of users, that motivation goal tends to be lost in the maze of bureaucratic labyrinth. The authoritarian environments (especially bureaucracies) suffer from a high risk of corruption, which dramatically reduces efficiency. This leads to a centralized business model that is very difficult to efficiently scale. The environment of responsibility empowers users to exercise their free will, but still makes them accountable for the consequence of their actions. It attempts to find the equilibrium between colliding forces and keep the system in that equilibrium by self-balancing mechanisms. We seek to design an architecture that will support the environment of responsibility.

The proposed architecture must respect privacy of users. We adopt the proposals of Solove [29] arguing that architectural approach is needed to remedy the problem of privacy in the information age. He proposes architecture based on Fair Information Practices [30] which was a one of the sources for OECD guidelines for the protection of privacy and which are also similar to rules defined by EU personal data protection directive [31].

An appropriate level of trust is necessary for efficient cooperation. The cooperation of the parties that do not trust each other is burdened with high overhead of contractual constraints, management controls and continual checks. The lower level of information about the trustworthiness of the other party means higher risk for the transaction and therefore higher cost to the controlling and remediation mechanisms rather than investing to the core subject of the interaction. We consider reputation as a key mechanism how to evaluate trustworthiness of the subjects. Interaction partners may user reputation values to guide trust-related decisions about the other partner without requirement for a long-term relationship. Such a method will be necessary to maintain efficiency in globally distributed environment.

5 Architecture of the Internet

The development of Internet is an effort of a large and diverse group. Computer scientists, application developers, users and many others, all are taking their part in forming the Internet. The development of protocols and applications is not controlled by any central authority. Several standard bodies are strongly influencing the design of the Internet, but even those organizations are not a position of force and control. Such environment contributed to the evolutionary approach to the architecture of the Internet. There is no document that describes the invariable architectural principles of the Internet, as there are no such principles [46]. As the design of Internet is guided by evolution and constant change, the occurrence of design problems and architectural inconsistencies is inevitable.

World Wide Web originated in early 1990s as an distributed hyper-text system, based on TCP-based Hyper Text Transfer Protocol (HTTP) and SGML-based Hyper Text Markup Language (HTML). It later evolved to a generic delivery mechanism for information objects. At the time of this writing is World Wide Web perceived as a general-purpose “information space” [41]. The World Wide Web architecture was guided by the Representational State Transfer (REST) architectural style described by Fielding [40]. The REST style is based on the Client-Cache-Stateless-Server style. All interactions are asymmetric, with the roles of client and server clearly distinguished. The server is passive (reactive) and cannot initiate interaction. All

interaction are limited to only those initiated by client, therefore asynchronous notifications of events from server to clients is not possible. This limitation is in practice addressed by polling mechanisms, such as RSS [47] and AJAX [48]. The server component of the REST architecture is supposed to be stateless [40]. The statelessness is endorsed as one of the key architectural principles of REST. When applied to the architecture of World Wide Web, the use of any state mechanism such as HTTP Cookies and frames is considered an architecture mismatch. However the assumption of statelessness can hold only if the resources are immutable and fail if any of them can be modified. The REST architecture allows for modification of resources, especially when applied to the WWW in a form of PUT, POST and DELETE methods of the HTTP protocol [39]. If one of the interactions changes state of the resource, all subsequent interactions depend on the result of the interaction that caused the state change. For that reason the interactions in the REST architecture cannot be considered stateless, as the state is present in the resources. Fielding does not address this problem in his description of REST, however he notes that the use of caching for resource representations may provide erroneous response. Such an error would not be possible if REST would be entirely stateless, in other words if the response would depend only on the information in request.

Uniform interface is another basic principle of REST architecture. However it was only partially reflected to the architecture of World Wide Web. The URI [38], HTTP [39] and HTML [49] specifications were supposed to define the uniform REST-like interface for the World Wide Web. However these definitions include a considerable degree of extensibility of the definition, focusing on the syntax of the interface and defining only the minimal semantic meaning when needed. While this approach allows to use the WWW mechanism for a broad range of applications, the definitions provided in URI and HTTP specifications are closer to definition of a network layer rather than application interface.

The concepts of Resource and Uniform Resource Identifier (URI) are central concepts in the architecture of the World Wide Web. However only vague definitions of a resource are available [41]. It is obvious that *resource* may be a realspace object and that resources are identified by URIs. That implies that one of the intents of the WWW architecture is to identify realspace objects by URIs. The World Wide Web architecture document [41] mentions the concept of URI owners and it recommends a good practice for URI owners to provide representation of the resource. It follows that realspace objects should have representation in cyberspace maintained by the owner of the URI. Such a representation is always subjective. There is no assurance that the URI owner is also the owner of the realspace "resource", therefore the representation of the "resource" provided by the URI owner can be harmful. This problem was recognized [50] and a solution was proposed by the W3C Technical Architecture Group [51] by not allowing to provide a representation of a resource that is not an information resource. Although it is claimed in the decision leaves a consistent architecture, some issues still remain. The most obvious problem is that the above decision makes generic concept of URI dependent on the HTTP protocol definition. However URIs are supposed to be protocol independent identifiers [38].

The definition of a resource is very vague. It is essentially defined as "whatever might be identified by a URI" [41]. This may lead to almost anything to be considered a resource. Even resource representations may by themselves be resources (they are often identified by URIs already). Such a recursive principle gives great freedom of choice for system implementers, but it may become very confusing. The situation may be easy to resolve for humans, given a specific context of the URI in the message and the response. But if an automated reputation

system would interpret the negative feedback of a user on a specific URI, it would be difficult to distinguish whether the image processing algorithm, photographer's skill, model's look or model's personality was meant as the target of the opinion in case of rating a resource representing itself as a digital photograph. This situation is made even more confusing by W3C recommending to avoid arbitrary URI aliases [41] for the same resource while at the same time recommending different URIs for something that can easily be considered different representations of a single resource [52].

The World Wide Web Architecture [41] document proposes a practice to avoid URI aliases. However URI aliasing is a common practice on the web today. It is a common practice that several URIs identify the same resource by using different capitalization, omitting slash characters or using different URI schemes. This practice is clearly in conflict with the practice proposed by World Wide Web Architecture document [41], however it is deemed acceptable by at least some members of W3C TAG [53]. We consider the practice of using URIs in different schemes to identify the same resource as harmful. The identification of access method should be determined by the client, it should not be a part of resource identifier. The eXtensible Markup Language (XML) [54] used for data representation on the World Wide Web introduced the concept of namespaces. The XML namespace mechanism is used for identification other resources as well, for example for identification of services. The name in a XML namespace is called Qualified Name (QName) and it is composed from URI-formatted namespace name and free-form local part. QNames are not URIs, however the World Wide Web Architecture document [41] strongly recommends the use of URIs for resource identification. Although the same document mandated mapping between QName and URI this practice is seldom followed, as there is no universal or recommended mechanism. We see this duality in using QNames and URIs to identify the same concepts (resources) as harmful to the architecture of World Wide Web. We account the difficulties in mapping between QNames and URIs to the unnecessary flexibility of generic URI format, which inhibits the attempts to design an universal mapping mechanism.

The URIs using the *http* scheme are considered by World Wide Web Consortium (W3C) Technical Architecture Group (TAG) to support persistence. The draft finding of the TAG [53] claims that URIs with *http* scheme support persistence as well as it is practically possible. However, the persistence of URIs with *http* scheme (HTTP URIs) depends on assignment of DNS name. DNS name assignment can be made reasonably persistent in the mid-term scope (few years) for well-established organizations. However it is difficult for individuals to obtain a DNS domain under their control. Therefore it is difficult to implement persistence for HTTP URIs scheme for individual users.

The W3C TAG draft finding on the use of metadata in URIs [55] recommends that URIs should be easy to understand, which is also supported by another document [56]. This practice may be interpreted to encourage the use of human-readable names in URIs. However, human readable-names are often subject to change, therefore such a practice inhibits persistence of URIs. The URIs with *http* scheme can support practical mid-term persistence for well-established organizations and hosting scenarios. But considering the current situation of DNS name assignment practice the use of HTTP URIs as general-purpose persistent identifiers is not practical.

The HTTP URIs cannot be considered pure identifiers, as they leak several implementation-specific details. They define the access protocol to use. Although it argued by W3C TAG [41]

that the `http` scheme prefix should not be understood as definition of access protocols, the practice of distinguishing the access protocol from URI prefix is considered acceptable by the document published by the same organization [53]. The specification of URI [38] states that there is a distinction between URI and URL, but it fails to define a method to distinguish them. The specification of HTTP URIs [39] does not provide such mechanism either. Considering a practice common in the Internet today and the architectural inconsistencies stated above, we must consider HTTP URIs to be addresses for a specific use with the HTTP protocol and not a generic identifiers.

Current architecture of World Wide Web assumes that information always comes from its authoritative source or a trusted proxy. The HTTPS mechanism is designed to be effective for protection of information under such assumption. However, the usability of HTTPS is limited when a parading of the "static Web" do longer apply. For example if a massive replication and data migration mechanisms are used, there is no single place of data transmission. The requested information may come from any node in the network that has a replica of that information. There is no single source of data transmission and there are no trusted proxies.

According to the principles of WWW architecture, any resource of relevance should be given an URI. The users of Internet can be seen as resources and they are definitely resources or relevance, therefore they should be given URIs. However, such practice is seldom used and there is no direct support for that in the World Wide Web standards or architecture.

The semantic web [57] is a proposed concept that builds on top of World Wide Web principles. The goal of the semantic web is not a distribution and hyperlinking of human-readable documents, but it is rather focused on the computer-processable description of objects. The objects are supposed to be described in XML-based data languages, such as RDF [58]. The semantic web object descriptions are supposed to be ordinary WWW documents accessible using WWW protocols (usually HTTP). The semantic web does not store realspace objects. A software system cannot store an apple or a car. It can only store information about the object (object description). The problems related to this subtle difference were already identified by Berners-Lee [50]. It may also be an incomplete claim that semantic web stores the cyberspace objects, as the semantic web itself may only reference them and the objects themselves could be obtained from other systems (using non-WWW protocols). The semantic web is still under development and it is not yet widely deployed. The opponents [59] of the semantic web concept describe severe obstacles to the feasibility and practicality of the semantic web deployment. Most described problems are caused by the unreliable data in the semantic web. We consider the described problems as a consequence of the subjectivity of crossing the realspace-cyberspace boundary. We argue that the same problems apply to the conventional World Wide Web. However the human consumers of World Wide Web can judge the reliability of the content, while computers cannot.

6 Proposed Architecture

Proposed architecture is based on a new architectural style RRSS, which is heavily inspired by REST style [40]. The style is combined with additional styles and constraints to form proposed architecture of World Wide Web.

6.1 RRSS Architectural Style

A new architectural style, RRSS, is proposed to improve architecture of World Wide Web. The style defines basic constrains of structuring and representing information in cyberspace,

while taking into consideration that the origins of these information may be in realspace and also that it may be used by realspace entities. The RRSS is a conceptual architectural style. The RRSS architectural style defines four basic elements: Resource, Representation, Source and Semantics. It also defines the relations and constrains for those elements. The RRSS architectural style derives directly from the empty (NULL) style.

Similarly to the authors of REST architectural style and the World Wide Web Architecture we find it is difficult to explicitly define a resource, as it can represent many different types objects and concepts. However unlike the authors of WWW architecture we want to be more specific about the meaning of the resource. Therefore we will define the resource indirectly by defining the properties of the resource:

Resource is a cyberspace entity.

Resource can represent both realspace and cyberspace object or concept.

Resource is a **complete, self-contained and consistent** representation of object or concept.

Resource state is **dynamic and volatile**. It cannot be assumed that a resource follows any specific state-transition model, unless it is explicitly constrained by additional information.

The resource could be used for a variety of purposes in many different applications. It is not known whether it is possible to design a single mechanism to represent any resource in a form suitable for all current and future applications. Therefore we explicitly support polymorphism by allowing broad range of resource representations.

Resource can be represented in the cyberspace by any number of **resource representations**.

Resource representation is a cyberspace entity. If a resource representation is complete, self-contained and consistent, it may be considered a resource. This possibility is introducing a recursion in the model which is consistent with the proposed architecture. The provision to allow any number of representation may be a concern for interoperability, unless a small set of mandatory representation formats is defined. This mandatory set may change slowly over time as the system evolves, but still assure interoperability of most of the software components. There we recommend:

Small number of **well-defined and stable resource representation formats** should be standardized and mandated. Resource should provide at least one representation in the standard format.

The information that forms a resource representing realspace concept is crossing the realspace-cyberspace boundary and therefore it must be considered subjective. We propose to generalize that principle and apply it to all resources:

Resource is always **subjective**.

This generalization would allow to hide the detail whether the resource source is in realspace or cyberspace from the users of the resource. As the resource is always subjective, a source of the resource must be considered to evaluate the qualities of the resource:

The identification of a **resource source** is an integral part of the resource.

According to the model introduced in chapter 3, resource source is a persona. It is not required that the resource source is a persona that represents a realspace person. Resource

source may be a persona representing a computer system, for example if the resource represents results of automatic summation of database statistics. However, we expect that many resource sources will represent realspace persons, as resources produced by such sources will be most meaningful and useful for the users. The resource source persona may not represent the realspace identity of the person. It may be a pseudonym or a persona that only partially reveals realspace information about the person. Such approach may allow to dynamically tune the trade-off between privacy and revealing of information for the purpose of inducing trust in the consumers of the information.

The meaning of the resource may not be interpreted properly having just the resource representation. For example user that sees a picture on his screen cannot be sure whether the resource represents the picture or the object shown on the picture. Therefore an unambiguous semantic description of the resource is needed for a clear understanding of the resource meaning:

> Resource can be semantically described in a standardized computer-readable form. The s**emantic description** is mandatory for all resources and it must be a part of any interface that is used to access the resource.

The basic concepts guiding the use of resources in cyberspace should form a foundation for any of its implementation. Therefore these concepts should not depend on any implementation-specific detail. For example the resources and representations should not depend on any specific communication mechanisms, protocol or interface:

> The concepts of resource, resource representations, resource source and resource semantic description must be implementation-independent.

The implementation mechanisms should be build on top of the basic concepts specified above. The implementation should depend on the basic concepts, not vice versa. The implementation must provide all the basic concepts and keep the proper relations between them. This approach will allow to modify or replace the implementation without changing the basic principles. It is relatively easy to adapt applications to new communication mechanism, while it is very difficult to adapt applications when the basic operational principles change.

The RRSS architectural style does not contain any constraint concerning resource identification. This omission is deliberate, as we consider resource identification as non-essential part of the RRSS architectural style. Any implementation of the style may provide their own means for distinguishing the resource, while some implementation may use hidden or implicit identification mechanisms (such as memory pointers). Therefore we will describe the aspects of resource identification as a separate architectural constraint that can be optionally applied to the RRSS architectural style. The resource identification constraint is defined as follows:

> Each resource has assigned an **identifier** that can uniquely and consistently identify the resource within the whole system.

For the purpose the reliable operation of the World Wide Web the identifiers must be unambiguously identify the resources. This requirement is partially expressed by the Resource Identification constraint defined in previous sections, mandating that the identifier uniquely and consistently identifies the resource. This can be further specified in a form of more concrete constraints:

> Resource Identifier must identify at most one resource.

Resource identifier that was assigned to a resource cannot be assigned to a different resource.

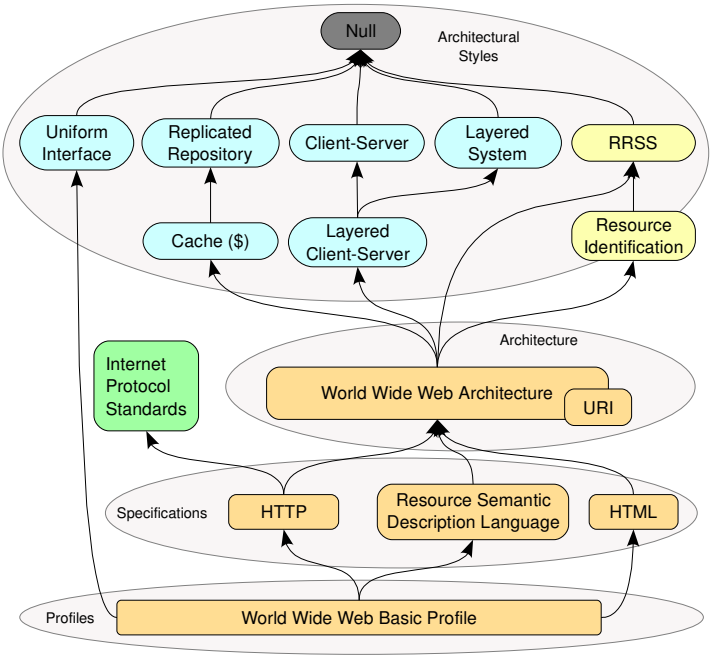
Resource identifier scheme must not depend on any structure or concept that underlies the implementation.

The semantics of the identifier must be opaque to the client applications.

6.2 Revised World Wide Web Architecture

We propose to split the overall World Wide Web architecture to several levels of abstraction. The split may improve the understanding and visibility to the architectural concepts. Proper layering of the abstraction can also address different goals of dynamics and interoperability properties of the architecture, as explained below. We propose following four levels of abstraction:

- **Architectural Styles** are the most abstract concepts. There form a set of architectural constrains that guide the creation of systems with appropriate properties and qualities. This layer consists of RRSS style and Resource Identification constraint described above, Layered Client Server, Cache and Uniform Interface styles described in [40].
- **World Wide Web Architecture** is a set of architectural constraints, rules and recommendations that define basic principles of World Wide Web operation. These principles are considered fundamental and it is expected that they will be valid and applicable for a long time. This layer is described below.
- **Protocol Specifications** provide specific definitions of communication protocols, data formats and interfaces. These specifications are based on basic principles, constraining them by specification of implementation details. It is expected that the protocol specification will be continually adapted to the implementation needs and that several protocols may exist at the same time for the same purpose, with different characteristics. This layer is described only marginally in this work.
- **World Wide Web Profiles** define a set of protocols that are required for correct cooperation of all World Wide Web components. Profiles are mechanism for interoperability. Profiles layer is described only marginally in this work.



World Wide Web Architecture layer is a composition of the architectural styles from the Architectural Styles layer supplemented by more specific architectural constraints. The goal of World Wide Web Architecture is to provide guidelines for developers of specifications that govern the basic operation of World Wide Web. It also specifies fundamental concepts of the World Wide Web, such as Unified Resource Identifier (URI). Figure XXX illustrates the composition of World Wide Web architecture as well as the abstraction layers that derive from it.

The constraint that the resources must be uniquely and consistently identified is a one of the fundamentals of World Wide Web simple hyperlinking scheme. One document (resource representation) may refer to another resource using the identifier. We define Uniform Resource Identifier concept to fulfill such role:

> Resources are identified by identifiers with fixed syntactic rules, called **Uniform Resource Identifier** (URI).

The complete syntax and semantics of Uniform Resource Identifiers should be part of the World Wide Web Architecture. The use of existing naming schemes for URIs should be deprecated and a single well-defined naming scheme should be defined. The naming scheme should conform to the Resource Identification architectural constraints defined above. All entities conforming to current URI syntax and not following the new URI scheme should be considered Uniform Resource Locators instead.

There is no constraint that would limit the number of identifiers assigned to a resource. Therefore different identifiers identifying the same resource are possible. Such identifiers are sometime called aliases. The existence of aliases makes it impossible to evaluate the equality of resources by comparing their identifiers. This situation may be confusing for applications, as applications may refer to the same resource without knowing that it is in fact the same resource. To address this problem we introduce the concept of canonical identifier:

> Each resource must be assigned exactly one **canonical Uniform Resource Identifier** at any specific time.

Canonical URI can be used as normal URI to refer to the resource. There can be only one canonical identifier per resource and each identifier may identify only a single resource. Therefore it can be easily evaluated whether two identifiers refer to the same resource by resolving the identifiers to canonical identifiers and comparing resulting canonical identifiers.

The World Wide Web Architecture layer further constraints the concept of resource presentation into a form that can be used in protocol and data format definitions:

> **Resource representation** is an array of bytes that form the content of the representation together with metadata that describe representation data format and may describe other aspects of the resource representation.

Based on the client-server architectural style, we define that resource representation is a data unit transferred between client and server:

> Resource representations can be directly retrieved from server to clients by using **access protocol**.

Resources are identified by URI. However the URI should provide no information about the location of the resource or its representations. We also place no constraints regarding location of resources and resource representations, not even the constraint that the representations of one resource should be placed together. Therefore we consider it useful to define a concept that will represent address of resource representations:

> Resource representations are addressed by **resource representation locator**.

The RRSS architectural style mandates semantic description of the resource. Our proposal of WWW constrains that and defines semantic description of a resource to be resource representation. Therefore it becomes a mandatory representation of a resource:

> **Semantic description** of resource is a mandatory representation of that resource.

There must be a clean distinction between resource metadata and representation metadata. Our solution is to make resource metadata a mandatory representation of the resource. By constraining semantic description of a resource to the limits imposed by resource representation we are simplifying the architecture. No additional special mechanism is needed to handle semantic descriptions. Ordinary access protocol used for resource representations can be used. This approach also allows the existence of several data formats for resource semantic description, for example to address innovation (migration from one format to another) and experimental usage. However we expect that the World Wide Web Profiles will mandate single data format for resource semantic descriptions.

Based on the model introduced in chapter 3, we can consider resource source to be a persona. As persona is a cyberspace entity and it can be presented using a set of cyberspace representation, we will constraint it the concept of a resource source to be a resource:

Resource source is a resource.

By mandating that resource source is a resource we solve several problems:

- The resource sources can be identifier in the same way as resources, using URIs.
- The representation of resource source can be accessed using the same mechanism as the resource itself, supporting uniformity and simplicity of the system.
- The mandatory representation of resource source is a semantic description, specifying the nature of the source.

According to the constraints of RRSS architectural style the identification of resource source is an integral part of the resource. Therefore a proper place in the architecture is needed to express the relation between resource and resource source. We consider the semantic description of the resource as an ideal place for that, as it is a mandatory element as its purpose is well aligned with the purpose of resource source identification. Therefore we mandate:

Identification of resource source should be mandatory element in the resource semantic description.

The global nature of URIs may endanger privacy of personas, if used incorrectly. If a single global identifier for a persona is used, then all visible actions of the persona and all accessible resource published by the persona may be correlated and an attacker may gain more information that was consciously released by the controller of persona. A decent level of privacy may be maintained if pseudonyms are used instead of primary identifiers. A different pseudonym for the same persona may be used for each site. Therefore the amount of information available to the attacker is limited and the danger of exposing additional information is lower. A simple approach to address this problem would be to use different URIs that identity the same resource. However, such approach will be hindered by the mechanism for evaluating URI equivalence. Such URI pseudonyms can easily be correlated using canonical URI. One possible solution is to create several resources that represent the persona. We can take advantage of the recursiveness of resource representation. We can model the original persona as a resource, while the pseudonyms will be representations of that resource. However, the pseudonyms are resources by themselves, therefore they have their own canonical URIs. The canonical URIs of the pseudonyms are different, therefore the pseudonyms are not trivially linkable.

6.3 Validation of the Architecture

Proposed architecture resides mostly on conceptual level. It is considerably abstract, with only a few concrete proposals. Therefore usual architectural validation by prototyping key elements is not applicable for this proposal. The architecture is not defined in a sufficiently formal way to mount a formal proof of correctness. Therefore we have decided to use scenario-based validation of the architecture. The scenarios used for demonstrating efficiencies of current World Wide Web architecture were used for validation of proposed architecture. The solutions to the scenarios compliant with proposed architecture are provided for each of the scenarios and the solution properties and variants are discussed. Each solution also contain a list of essential architectural elements used in the solution.

Although the scenario-based validation cannot prove formal correctness of the architecture, it increases confidence in the architecture appropriateness and usability in a scenarios that are

close to practice. It is also used a checking mechanism to uncover any obvious problems. The list of essential architectural elements in each scenario is used to make sure that all of proposed architectural constructs are used and therefore none of them is obviously redundant.

The goal of the validation is to show that the proposed architecture is an *improvement* over the original World Wide Web architecture, which is demonstrated on the scenarios described in the dissertation.

Proposed architecture introduces a couple of drawbacks and trade-offs. Multi-step name resolution is a direct consequence of URI abstractness. Location-independent identifiers naturally introduce additional level of indirection therefore increasing the overhead of name resolution. Mandating semantic description and resource source identification introduces additional complexity to web servers. However web applications are usually already aware of the nature of resources they present and they usually also know the source of the resources, therefore it may be feasible to handle such additional complexity.

The proposed architecture is specified on conceptual level only. Although the architecture was validated by walking through a set of scenarios, it is expected that more drawbacks will surface when the architecture will be applied to the design of individual components, protocol specifications and profiles. We do not consider such event to be a failure of proposed architecture as long basic architectural principles holds. Such problems are expected and once they are uncovered they should be fed back to the architectural process, inducing changes in the proposed architecture, following a sound iterative development approach.

7 Contributions to the Field

The motivation of this work was to discuss what went wrong in the course of Internet evolution and to propose and improvements that could help change the Internet to a more desirable environment. The focus of this work was on the most important part of the Internet: the users. To lay the foundation for this work, we have examined the interactions of the physical personas with computer systems. In chapter 3 we have proposed a model that can explain some of the aspects of such interactions. It was demonstrated how the proposed model could be used to evaluate such properties as identity and anonymity. We consider the model still being in its infancy and we expect that follow-up work will substantially extend and improve it. However, we have demonstrated the usefulness of the model by applying it to the evaluation of current World Wide Web architecture, identification of architectural problems and proposals of improvements.

This dissertation have identified major problems in the Internet architecture, especially the problems of the architecture of World Wide Web. The major problem areas include:

- Addresses being used for identification purposes, including IP address and some URIs.
- Assumption that World Wide Web resources are static.
- Weak definition of World Wide Web interfaces and inconsistent application of architectural principles.
- Vague meaning of World Wide Web resources.
- Protocol-dependence of World Wide Web Architecture.

- Inappropriate security mechanisms.

Revised architecture of World Wide Web is proposed in chapter 6. The proposal is based on the evaluation of the problems of current World Wide Web architecture. It is attempt to amend the basic architectural principles and guidelines, especially in a way that would reflect the implications of the model provided in chapter 3. We have proposed to divide the field into several four layers of abstraction with different requirements for design stability and dynamics.

The architectural styles layer defines RRSS, a new architectural style inspired by the REST style. The RRSS style builds on four basic components: Resource, Representation, Source and Semantics. While the concepts of resource and resource representations are adopted from the REST architectural style, the concepts of resource source and semantic description are our additions to the architectural style. Especially the concept of resource source is an implication of our model on the architecture of World Wide Web.

The RRSS architectural style, combined with other styles is applied to the World Wide Web architecture. The result is (still abstract) set of architectural guidelines that govern the basic principles of World Wide Web and provide foundation for protocol specifications. The World Wide Web architecture combines the architectural constraints of RRSS and other styles to adapt them for the World Wide Web environment. A redefined concept of Uniform Resource Identifier (URI) is outlined in a form of ideas and requirements, while specific definition of the URI is left for future work. The concept of resource source is specified in a more concrete form. It is defined as a resource representing a persona of resource owner. Such recursiveness simplifies the principles of the model, while not constraining its flexibility. Security and privacy aspects of proposed architecture are shortly discussed as well.

The basic ideas of World Wide Web specifications and profiles are outlined, however they are not considered to be a focal point of this work. The end of chapter 6 summarizes a set of changes to current World Wide Web architecture document to make it compliant with proposed architecture. A short outline of a migration process from current World Wide Web to the proposed architecture is also provided.

The primary principle of this work is that the crossing of realspace-cyberspace boundary is subjective. Therefore all information in the cyberspace should be regarded as *opinions*, rather than unquestionable truths. The implication of that principle is the introduction of the concept of *source* to the RRSS architectural style. This concept is then reflected to the World Wide Web architecture in a form of resource source identifier in a resource semantic description. Such an approach will effectively make the identity-related mechanisms an integral part of World Wide Web architecture.

The objectives of the dissertation that were defined in chapter 2 were met. This work makes the following contributions to the research within the field of Information and Computer Science:

- A model that describes interactions between realspace and cyberspace, especially focused on representation of personal data in the cyberspace.
- A mechanism for evaluation of anonymity and identity based on the proposed model.

- Assessment of architectural inconsistencies of World Wide Web architecture. Both internal inconsistencies and problems uncovered by the application of the proposed model are described.
- A definition of RRSS, a new architectural style for representing information in cyberspace while taking into consideration their potential source and target in realspace.
- Application of the RRSS architectural style together with other previously described styles to a World Wide Web architecture, resulting in a proposal of improved architecture for World Wide Web.

Conclusion

The Internet and World Wide Web are revolutionary technologies. They allow people to cooperate and efficiently share information. The Internet was not created in its current form, it has rather evolved in time. Similar evolution also applied to the World Wide Web, however that was partially guided by the architectural principles of REST. Such evolutionary approach worked perfectly to address simple needs of the environment and guarantee the survival of the Internet and World Wide Web. However, it failed to address more complex needs, such as privacy and data authenticity.

Up until recently most of the information available on the World Wide Web were public or intended for public usage. This paradigm of the "static Internet" is changing. Large amount of non-public information is being transferred over the Internet. This information cannot be simply classified as private or public. The classification is of much finer grain and contains a degree of fuzziness: share information with my friends, with magazine subscribers, with business partners, with premium customers, etc. The organization of the Internet into strictly separate sites may also be challenged. The rise of peer-to-peer networks does not operate on the concept of a site. Replication and migration of data on demand is a modus operandi of these networks. Therefore the concept of "source of data transmission" is no longer useful for these networks. The parading of "few publishers, many readers" has transformed to "few publishers, many contributors, hordes of readers". The World Wide Web is no longer read-only information system, it becomes writable. However, these paradigm shifts are not well supported by current World Wide Web architecture.

We have proposed an improved architecture of the World Wide Web, adapting the basic architectural concepts to meet new requirements. Our architectural work was guided by a model of realspace-cyberspace interactions, that guided basic architectural ideas. The architectural form was shaped according to its desired function, seeking to induce good cooperation in the environment of responsibility and preserve privacy of the users.

The work to reshape the World Wide Web is far from complete, it is rather at its very beginning. This work lays a basic conceptual foundation for future works that can add more technical details, shape the specification and test the system in practice. The architecture of Internet-based system cannot grow on a green field. They need to coexist with currently deployed and widely used technologies, even if the deployed technologies are far from ideal. Our architectural proposal is formed as an extension and improvement of existing system: World Wide Web itself. We hope and believe that our work can help change the thinking behind the World Wide Web architectural to be more focused on the nature of the provided

information, more supporting to the ad-hoc cooperation of people while preserving their privacy.

Zhrnutie

Internet a World Wide Web boli vyvíjané evolučným spôsobom nasledujúc meniace sa požiadavky používateľov a spôsoby využívania technológií. Momentálne požiadavky na dynamiku informácií, interaktivitu a efektivitu spolupráce nútia technológie k ďalšej zmene. Táto zmena zasahuje hlboko do základných princípov a preto je potrebné revidovať jadro architektúry World Wide Web-u.

Cieľom práce je zlepšiť architektú WWW s ohľadom na zmenené požiadavky. Práca sa zameriava na konceptuálnu úroveň architektúry, keďže nie je reálne v jedinej práci obsiahnuť tak širokú tému ako je kompletná architektúra WWW.

Práca predstavuje model popisujúci interakcie medzi skutočným svetom a virtuálnym svetom ktorého cieľom je lepšie porozumenie základným princípom a obmedzeniam. Model definuje dva svety: reálny svet (*realspace*) a virtuálny svet (*cyberspace*). Informácia, ktorá prechádza medzi týmito dvoma svetmi musí byť považovaná za subjektívnu, keďže spoľahlivosť takejto informácie nie je možné priamo overiť. Zdroj z ktorého informácia pochádza je rovnako dôležitý ako informácia sama. Model je zameraný najmä na informácie o reálnych osobách vo virtuálnom svete. Dátová štruktúra, ktoré opisuje reálnu osobu je nazvaná persóna. Model skúma vzťahy medzi osobami a persónami z ktorých vyplýva anonymita alebo identita osoby. Hranica reálneho a virtuálneho sveta však zabraňuje použitiu týchto pojmov v praxi, preto model definuje pojmy analógie a heterológie persón. Analógie a heterológie je možné vyhodnotiť priamo na persónach a preto môže byť použitá ako praktická aproximácia identity a anonymity.

Správna architektúra musí vychádzať z vlastností a obmedzení prostredia ale najmä z požiadaviek používateľov a cieľov ktoré má naplniť. Z tohto dôvodu práca diskutuje o žiadúcom prostredí ktoré má architektúra podporovať. Práca diskutuje o požiadavkách na súkromie používateľov, zameraná najmä na vhodný prístup k spracovaniu osobných údajov. Dôvera medzi používateľmi Internetu je rozoberaná, najmä s ohľadom na ad-hoc interakcie medzi používateľmi bez požiadaviek na predchádzajúci dlhodobý vzťah. Mechanizmy založené na princípoch reputácie používateľov sú doporučované ako možné riešenie.

Model a analýza prostredia sú použité na zhodnotenie aktuálnej architektúry World Wide Web-u. Práca odhaľuje nesúlad architektúry WWW aktuálnymi požiadavkami ako aj vnútorné architektonické inkonzistencie, ktoré sú v práci detailne popísané. Najmä nejasné definície a rozporuplné architektonické rozhodnutia sú považované za prekážky ďalšieho rozvoja WWW. Problémy sú demonštrované na sade scenárov, ktoré je zložité implementovať pri zachovaní súčasnej architektúry.

Práca navrhuje zlepšenú architektúru v podstatnej miere založenú na novom architektonickom štýle nazvanom RRSS. Tento štýl kombinuje štyri základné pojmy: informáciu (*resource*), jej reprezentáciu (*representation*), zdroj informácie (*source*) a sémantický popis (*semantics*). Všeobecný architektonický štýl RRSS je použitý pri návrhu zlepšenej architektúry WWW.

Navrhnutá architektúra je rozdelená na niekoľko úrovní abstrakcie pre ľahšie porozumenie a jednoduchšiu údržbu architektúry: úroveň architektonických štýlov, úroveň architektúry WWW,

úroveň špecifikácií a úroveň profilov. Úroveň architektonických štýlov je veľmi abstraktná, čo obmedzuje nutnosť častých zmien ale vylučuje jej priamu praktickú aplikáciu. Na druhej strane úroveň profilov je veľmi konkrétna a priamo implementovateľná, čo však znamená veľké množstvo zmien. Keďže práca je zameraná na konceptuálnu úroveň WWW, pozornosť bola venovaná najmä na úrovni architektonických štýlov s presahom do úrovne architektúry.

Navrhovaná architektúra je overená validáciou založenou na scenároch. Je ukázané ako navrhovaná architektúra dokáže podporovať scenáre, ktorých implementácia je zložitá pri súčasnej architektúre WWW.

Bibliography

[1] Schneier, B.: Applied Cryptography, Protocols, Algorithms, and Source Code in C, John Wiley & Sons, ISBN: 0-471-11709-9, 1996

[2] Haller, N.: The S/Key One-Time Password System, ISOC Symposium on Network and Distributed Systems, 1994

[3] Haller, N., Metz, C., Nesser, P., Straw, M.: A One-Time Password System, RFC 2289, 1998

[4] Menezes, A., Van Oorschot, P., Vanstone, S.: Handbook of Applied Cryptography, CRC Press, ISBN: 0849385237, 1996

[5] Simpson, W.: PPP Challenge Handshake Authentication Protocol (CHAP), RFC 1994, 1996

[6] Information Technology - Open Systems interconnection ..., ITU-T Recommendation X.509, 2000

[7] Dierks, T., Allen, C.: The TLS Protocol Version 1.0, RFC 2246, 1999

[8] Rescorla, E.: HTTP Over TLS, RFC 2818, 2000

[9] Hoepner, P. (Ed.): Study on PKI and Certificate Usage in Europe 2006, Fraunhofer Institute FOKUS, 2006

[10] Clarke, R.: The Fundamental Inadequacies of Conventional Public Key Infrastructure, Procceedings of ECIS'2001 conference, Bled, Slovenia, 2001

[11] Brands, S.: Rethinking Public Key Infrastructures and Digital Certificates, MIT Press, ISBN: 0-262-02491-8, 2000

[12] Liberty Bindings and Profiles Specification, Liberty, 2003

[13] Liberty Bindings and Profiles Specification, Liberty, 2003

[14] Assertions and Protocol for the OASIS Security Assertion Markup Language, OASIS, 2003

[15] Web Services Federation Language (WSFederation), BEA, IBM, Microsoft, RSA Security, Verisign, 2003

[16] WS-Federation: Passive Requestor Profile, BEA, IBM, Microsoft, RSA Security, Verisign, 2003

[17] Nanda, A.: Identity Selector Interoperability Profile V1.0, 2007, http://download.microsoft.com/download/1/1/a/11ac6505-e4c0-4e05-987c-6f1d31855cd2/

Identity-Selector-Interop-Profile-v1.pdf

[18] OpenID Authentication 2.0 - Final, , 2007, http://openid.net/specs/openid-authentication-2_0.html

[19] Cantor, S., et al.: Shibboleth Architecture, Protocols and Profiles, 2005, http://shibboleth.internet2.edu/shibboleth-documents.html

[20] The Simple eXtensible Identity Protocol (SXIP) Reference, , 2004, https://sxip.net/archive/specs/sxip-reference.pdf

[21] Light-Weight Identity, NetMesh Inc., 2005

[22] Pashalidis, A., Mitchell, C.: A taxonomy of single sign-on systems, Information Security and Privacy, ACISP 2003, 2003

[23] Beginner's guide to OpenID phishing, , , http://marcoslot.net/apps/openid/

[24] OpenID: Phishing Heaven, , 2007, http://www.links.org/?p=187

[25] Abelson, H., Lessig, L.: Digital Identity in Cyberspace, White Paper Submitted for 6.805/ Law of Cyberspace: Social Protocols, 1998

[26] Pfitzmann, A., Köhntopp, M.,: Anonymity, Unobservability, Pseudonymity, and Identity Management A Proposal for Terminology, Designing Privacy Enhancing Technologies, International Workshop on Design Issues in Anonymity and Unobservability, 2000

[27] Anonymity, Unlinkability, Undetectability, Unobservability,Pseudonymity, and Identity Management –A Consolidated Proposal for Terminology v0.31, , 2008, http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.31.pdf

[28] Warren, S., Brandeis, L.: The Right to Privacy, Harvard Law Review, Vol. IV, No. 5, 1890

[29] Solove, D.: The Digital Person: Technology and Privacy in the Information Age, NYU Press, ISBN: 0814798462, 2004

[30] Records, Computers and the Rights of Citizens, Report of the Secretary's Advisory Committee on Automated Personal Data Systems, U.S. Department of Housing, Education, and Welfare, 1973

[31] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, European Parliament and the Council of European Union, 1995

[32] Resnick, P., et al.: Reputation Systems, Communications of the ACM, 43(12), 2000

[33] Axelrod, R.: Evolution Of Cooperation, Basic Books, New York, ISBN: 0465021220, 1984

[34] Windley, P., et al.: Using reputation to augment explicit authorization, Proceedings of the 2007 ACM workshop on Digital identity management, 2007

[35] Sabater, J., Sierra, C.: Review on computational trust and reputation models, Artificial Intelligence Review, Vol. 24(1), 2005

[36] Friedman, E., Resnick, P.: The Social Cost of Cheap Pseudonyms, Journal of Economics and Management Strategy, Vol. 10, 2001

[37] Berners-Lee, T.: Information Management: A Proposal, 1990, http://www.w3.org/History/1989/proposal.html

[38] Berners-Lee, T., et al.: Uniform Resource Identifier (URI): Generic Syntax, RFC 3986, 2005

[39] Fielding, R., et al.: Hypertext Transfer Protocol -- HTTP/1.1, RFC 2616, 1999

[40] Fielding, R.: Architectural Styles and the Design of Network-based Software Architectures, University of California, Irvine, 2000

[41] Architecture of the World Wide Web, Volume One, W3C Recommendation, 2004, http://www.w3.org/TR/2004/REC-webarch-20041215/

[42] Dhamija, R., Tygar, J. D., Hearst, M.: Why Phishing Works, Proceedings of CHI 2006 Conference on Human Factorsin Computing Systems, 2006

[43] Internet Usage Statistics, The Internet Big Picture, , 2008, http://www.internetworldstats.com/stats.htm

[44] Shannon, C.E.: A Mathematical Theory of Communication, Bell System Technical Journal, ,

[45] Economides, N.: The Economics of Networks, Brazilian Electronic Journal of Economics, vol. 1(0), 1997

[46] Carpenter, B. (Editor): Architectural Principles of the Internet, RFC 1958, 1996

[47] RSS 2.0 Specification, , 2003, http://cyber.law.harvard.edu/rss/rss.html

[48] Garrett, J.: Ajax: A New Approach to Web Applications, 2005, http://www.adaptivepath.com/ideas/essays/archives/000385.php

[49] HTML 4.01 Specification, W3C Recommendation, 1999, http://www.w3.org/TR/html401/

[50] Berners-Lee, T.: What do HTTP URIs Identify?, 2002, http://www.w3.org/DesignIssues/HTTP-URI.html

[51] Berners-Lee, T.: What HTTP URIs Identify, 2005, http://www.w3.org/DesignIssues/HTTP-URI2.html

[52] Raman, T.V. (Editor): On Linking Alternative Representations To Enable Discovery And Publishing, 2006, http://www.w3.org/2001/tag/doc/alternatives-discovery.html

[53] Thompson, H.S., Orchard, D.: URNs, Namespaces and Registries, 2006, http://www.w3.org/2001/tag/doc/URNsAndRegistries-50

[54] Extensible Markup Language (XML) 1.1 (Second Edition), W3C Recommendation, 2006, http://www.w3.org/TR/xml11/

[55] Mendelsohn, H., Williams, S. (Editors): The use of Metadata in URIs, 2006, http://www.w3.org/2001/tag/doc/metaDataInURI-31-20061204.html

[56] Berners-Lee, T.: Cool URIs don't change, 1998, http://www.w3.org/Provider/Style/URI.html

[57] Berners-Lee, T., et al.: The Semantic Web, Scientific American Magazine, May 17, 2001

[58] Klyne, G., Carroll, J. (Editors): Resource Description Framework (RDF): Concepts and Abstract Syntax, , http://www.w3.org/TR/rdf-concepts/

[59] Doctorow, C.: Metacrap: Putting the torch to seven straw-men of the meta-utopia, , http://www.well.com/~doctorow/metacrap.htm