Article

International Federation of Library Associations and Institutions
2018, Vol. 44(3) 170–182

DOI: 10.1177/0340035218777275
journals.sagepub.com/home/ifl
SAGE

Privacy awareness issues in user data collection by digital libraries

Elaine Parra Affonso
São Paulo State University (UNESP); Faculdade de Tecnologia (FATEC), Brazil

Ricardo César Gonçalves Sant'Ana
São Paulo State University (UNESP), Brazil

Abstract

This work has the objective of investigating privacy aspects in the collection of data by the National Digital Libraries of South America. Country-specific digital libraries were examined using an exploratory research method to identify data these libraries collected both with the user's awareness and in the explicit presence of privacy policies within their environments. Brazil's National Digital Library environment was also examined by using the Wireshark tool to identify possible data collected implicitly during user interaction. We identified that only two of the examined digital libraries provide privacy guidance, and in relation to the collection process, the data that are collected without the knowledge of the user stand out more than the data that the user makes available consciously. It is concluded that privacy issues can be influenced by low user awareness of when, how and where data collection takes place, and the availability of privacy policies becomes essential in digital libraries to raise awareness about this process.

Keywords

Abstraction layers, awareness, data collection, privacy

Submitted: 18 September 2017; Accepted: 14 February 2018.

Introduction

With the increased use of technological devices, activities that realize data collection increase, reaching all segments of society. As such, it becomes necessary to better understand this process which often does not occur in a perceptible way to the user who has low awareness about when, how, and where it occurs. Since data relating to such actions may reveal individuals' personal information, threats to privacy emerge. Tanenbaum and Wetherall (2011) point out that due to rapid technological growth, the differences between data collection, storage, and processing are rapidly disappearing, making issues in this process intangible to the user.

The effect that information technology has on privacy causes new concerns and can be analyzed from four factors: the amount of information collected by digital devices and environments; the speed with which information can be shared; length of storage time; and the type of information that can be collected (Tavani, 2008).

Threats to privacy extend from the moment that the user transfers their activities to the digital medium and leaves traces of interactions in those environments. According to O'Hara and Shadbolt (2014) each time a new technology emerges that allows communication and interaction without the need for physical presence, a new level of abstraction is created, because as long as there is no physical presence, the individual leaves representations in the environment making it harder to hide their interactions.

In addition to the data collection performed in digital environments explicitly, there are data that

Corresponding author:
Elaine Parra Affonso, São Paulo State University, Hygino Muzzi Filho Avenue 737, Marília, São Paulo State, Brazil.
Email: elainepff@gmail.com

circulate silently in computer networks. Silent data circulation results in lack of awareness into the data collection process, causing informational asymmetry[1] between data holders and users. It is emphasized that information asymmetry provides more power for those who hold the data, especially when it comes to personal data, and increases the lack of control over the collection. According to Mayer-Schönberger (2011), the loss of control is rarely transparent to the user since it occurs without the individual perceiving. In this way, when the individual loses control, others gain in the power of information.

In the digital libraries scenario, data collection can occur at the moment of user interaction when performing a search or when filling in registers to request information – including data traffic in computer networks. These environments must provide measures and guidelines regarding data collection issues that can identify individuals. According to Klinefelter (2016), digital libraries, while providing free access to information, also imply new privacy risks. This form of access often requires the user to identify themselves and their own interaction with the environment that leaves digital traces sufficient enough to be used in the commercial environment or by government agencies (Klinefelter, 2016).

In libraries, privacy is essential as it allows the user to choose and access information without fears, judgments, or punishments. The right to read can be compromised if the individual's privacy is threatened, and true freedom of choice in libraries requires both a variety of materials and the assurance that interaction and choices are not being monitored (ALA, 2017).

This study aims to investigate the privacy aspects in the data collection phase using the National Digital Libraries of South America as a basis. The following questions guide this study:

1. What data are collected during user interaction with the digital library site?
2. Are the data that is collected perceptible to the user? Or does the very interface of this process diminish the perception about the data collected?
3. Does the collected data imply privacy threats?
4. Are there privacy policies that explain to the user what data are collected during interactions on digital library sites?

Methodology

The methodology used in this study was based on: (1) identification of National Digital Libraries of South American countries through the Google search engine with the term national digital library descriptor and country name; (2) exploratory research on digital library sites to identify the following issues: explicit provision of privacy policies; communication protocol used; identification of data collected with the user's awareness; (3) identification of possible data collected implicitly in the user's interaction with digital environments, specifically with Brazil's National Digital Library.

In order to identify and demonstrate the possible data implicitly collected, and the elements involved in the data collection phase during the user's interaction with Brazil's National Digital Library, the Wireshark[2] tool was used. Through the Wireshark tool, it is possible to analyze each data packet that the user received and sent to the destination, verifying the source IP and destination IP data, number of ports, date and time of the request. When the page does not use encryption, it is possible to check the data sent from the user to the digital environment.

In this study, the Wireshark version 2.4.0 was installed and executed in the author's own equipment so that it was possible to initiate the capture and identification of traffic data packets when searching in the digital collection of Brazil's National Digital Library website, which consisted typing the title of the book "o cortiço" in the search field. The packets trafficked during access to the website were collected and a package was selected for analysis and exemplification of possible collection performed by the data holders during the data collection phase.

Subsequently, the main data present in the package were correlated with the layers of the Open System Interconnection (OSI) model, including verification whether the data are identifiers, quasi-identifiers, or sensitive.

The collection of data on the website of Brazil's National Digital Library was carried out through a notebook with the Windows operating system, with the wireless connection. It should be emphasized that the collection performed during the user's interaction with the digital library website was done by the authors' equipment and using the home network. Only the data resulting from this interaction have been viewed and analyzed. The data collection of this research was carried out in July 2017.

This text is divided into the following sections: Collection phase and privacy issues; Open System Interconnection (OSI) model and abstraction in the collection phase; Tool for data collection in computer networks; Results and discussions, and Considerations.

The main contribution of this article is to highlight the phase of data collection in the web environment,

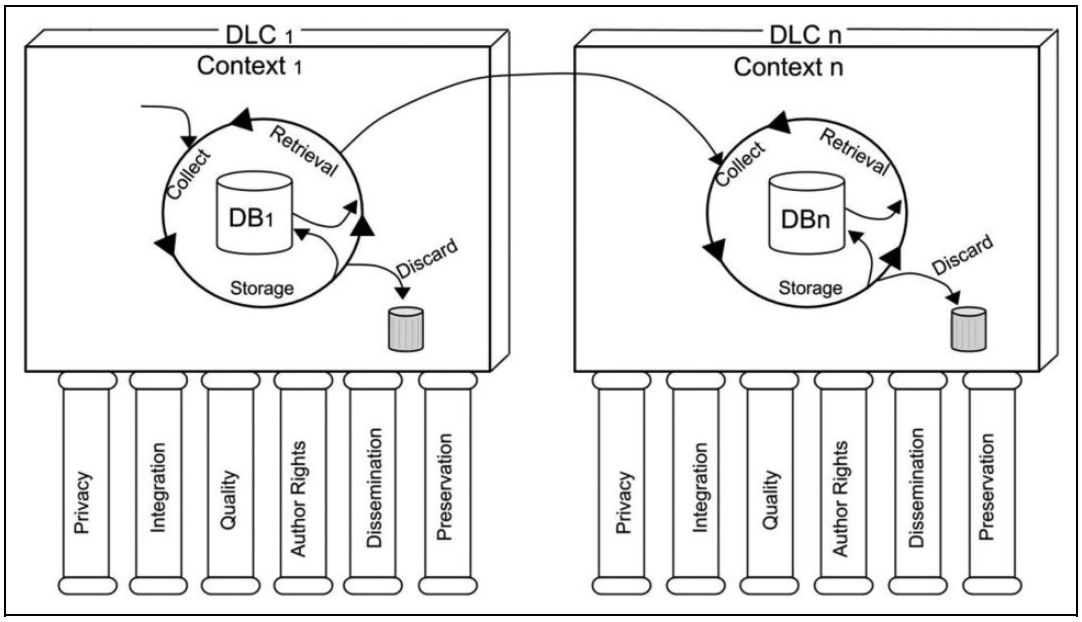

Figure 1. Data life cycle for information science.
Source: Sant'Ana (2016: 123)

explicitly in digital libraries, in order to demonstrate that the collection of data exceeds the data made available by the user, and the architecture of the communication networks themselves contributes to making this process more distant from the user. As a consequence, privacy threats increase.

Collection phase in the data life cycle and the privacy aspects

As a way to highlight the different moments and objectives present in the access and use of data, Sant'Ana (2013) proposes the Data Life Cycle for Information Science (DLC-IS). DLC-IS is a theoretical framework delimited in four phases: collection, storage, retrieval, and disposal. The phases are permeated by the factors privacy, integration, quality, author rights, dissemination, and preservation (Figure 1). This model seeks to contribute to a better understanding of these phases and involved resources.

The collection phase delimits the moment in which the purpose is to obtain data and in which the planning and execution of several activities occurs, among them: identification of the need for collection; definition of the data to be collected; procedures for collection; data format; and treatment necessary for the intended purpose of the collection (Sant'Ana, 2016). In the DLC, the collection phase is permeated by the factors privacy, integration, quality, copyright, dissemination, and preservation of data (Sant'Ana, 2016).

Among these factors, the collection of data can cause threats to the privacy of the individuals who participate in the collection context. In the case of this research, the user's privacy issues are taken into account in relation to the collection made by the data keeper, in the case of digital libraries. Regarding data quality, origin, collection, reliability, utility, and physical and logical integrity guarantees, these are fundamental at this stage of the data life cycle.

Regardless of the factors involved, digital environments such as digital libraries, social networks, search engines, mobile applications and the most diverse applications collect data with the justification of providing better results for users who make use of these environments. However, it is necessary to make users aware of the data collected and the privacy implications of individuals interacting with these environments.

The World Digital Library (2017) describes in its privacy policy that the environment offers a better service through the collection and storage of non-personally identifiable information and cookies. Their privacy policy also states that it only collects personal information the user voluntarily provides, and the use is intended only for the service. In addition, there is mention of the implementation of safeguards to protect any information collected.

The social network Facebook (2017) describes in its privacy policy that it collects data regarding the activities of the users and the information made available by said activities, including data about people and groups with which it connects. In addition to interactions and information, Facebook also collects data from payments, devices, sites and applications that use Facebook services, as well as information from external partners and companies of this social network.

Data collection can happen in two ways: directly involving the user, and collections that do not directly involve the user. When the user fills out a form on a website, he is aware that the collection is happening and understands that this activity brings benefits even though they do not understand the privacy implications. On the other hand, when browsers send cookie information back to the site or when surveillance cameras record activities in an environment, the collection occurs without user involvement (Spiekermann and Cranor, 2009).

Information that may seem harmless can be linked to new contexts., and it becomes difficult to get a sense of when privacy has been violated. As such, the Web becomes an environment that gathers more information about the user than other environments making it possible to construct an image of the user using the Web (Nissenbaum, 2011: 36).

When the user is interacting in a digital environment, a set of personal data is revealed to the data keeper. This personal data can be classified as: identifiers that uniquely identify the individual; quasi-identifiers that, when combined with other databases, allow the identification of the individual; sensitive data that reveal confidential information and, where disclosed, may place the data subject in situations of constraints; or non-sensitive data – the collection or dissemination of which does not imply privacy threats (Samarati, 2001).

Furthermore, according to the new Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and the free movement of such data, personal data may be defined as:

> information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. (GDPR, 2016: 33)

The problems related to the collection of user data and privacy are numerous and with consequences that are not yet estimated or perceived by individuals. Consequences may include issues of discrimination, induction in the choice of products and services, and correlation of data for the construction of user profiles. Fabian et al. (2010) point out that due to the repression imposed by some political regimes, in which copyright, freedom of expression and, in particular, free access to information are restricted, the various possibilities of data collection by various means can lead to the pursuit of individuals if their identity is revealed.

Through the dissemination of privacy policies, these environments are designed to offer users an awareness of data collection. However, the perception of the user may be linked to the description that the company makes available in these documents or in the data that the user makes available during the use of the service, such as username, passwords, field fields, search terms, among others. Thus, awareness about data collection involves the user's knowledge about how their data will be collected. The purposes of privacy policies should be to make information about data collection more clear and accessible and to broaden the user's perception of this process.

In these digital environments, computer networks are essential (specifically the Internet, attracting myriads of new users) and make it possible to configure several pages of information containing texts, figures, sounds, and video with embedded links to other pages (Tanenbaum, 2003).

To minimize the complexity involved in the operation of these communication networks, they are organized into layers of abstraction whose purpose is to provide services to the upper layers, isolating these layers from the implementation details (Tanenbaum and Wetherall, 2011). The concept of abstraction is common in computer science, receiving various names such as information hiding, abstract data types, and encapsulation (Tanenbaum and Wetherall, 2011).

In this way, the digital environments, when collecting data using computer networks, rely on a layered model of abstraction with the purpose of hiding from the user technical details of the activities and data collected. The most important abstraction principle in the field of communication in computer networks is the OSI reference model.

With this in mind, the layered approach proposed by the OSI reference model becomes relevant to hide details that are not operationally important to the user. Although it visualizes the communication process in the computer networks in a generalized way and with reduction of complexity, the layers specified in the model facilitate the understanding of moments and elements involved in this process, including data collection.

OSI reference model

The OSI reference model is a layered structure whose purpose is to inform the function of each layer and to keep software or hardware details hidden when

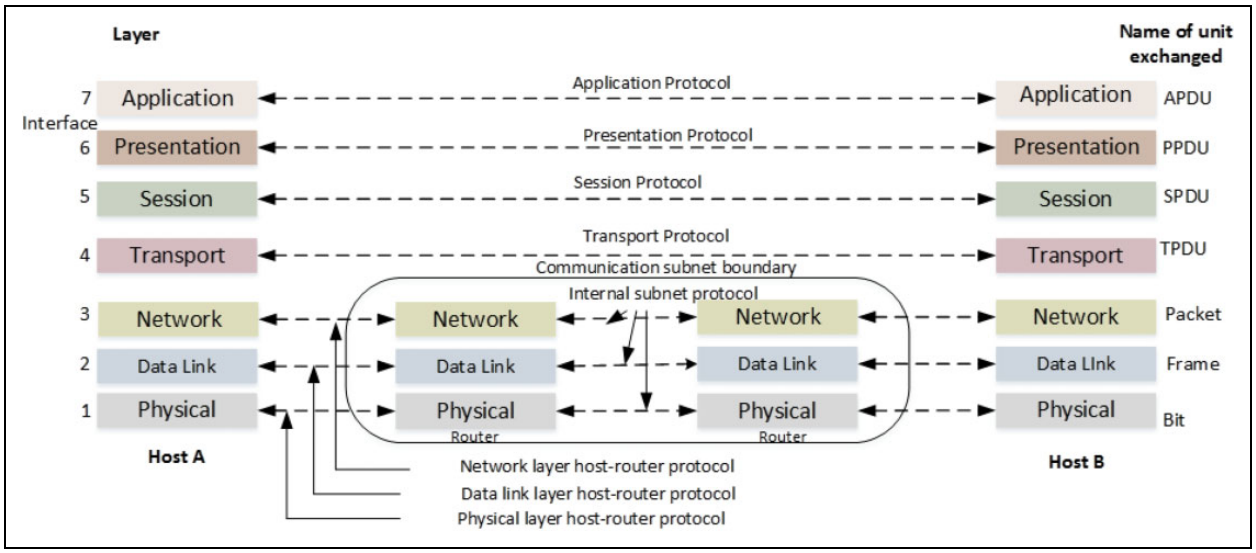

Figure 2. Layers of abstraction OSI reference model.
Source: Tanenbaum (2003: 41)

providing service to users (Tanenbaum and Wetherall, 2011). This model presents three concepts: services, interface, and protocol, making their differences explicit. The OSI model, through each layer, performs services for higher layers, which, in turn, determine what the layer accomplishes by defining the semantics of the layer. The interface informs how the upper layer processes can be accessed and the protocols make the work feasible, that is, they provide the services (Tanenbaum and Wetherall, 2011). This model was developed by the International Standard Organization (ISO) as a means to internationally standardize the protocols that are used in the layers: physical, link, network, transport, session, presentation and application (Tanenbaum, 2003) (Figure 2).

The OSI layered model helps to organize and simplify the understanding of operational concepts that might otherwise be unnecessarily detailed and complex, simplifying the complexity of computer network protocols and technologies by abstracting them from each other in multiple tiers (Nikkel, 2005).

Abstraction is fundamental in dealing with complexity. Its purpose in the network environment is to ignore small differences between the elements and processes of communication networks by considering only their similarities. An efficient abstraction is one that highlights important details for the user without considering those that are irrelevant to interaction (Sclavos et al., 1994).

The abstraction provided by the OSI model in communication networks can have an effect on the privacy of individuals interacting in the network environment by hiding different types of data collected during this interaction such as: the result of access to social networks, search engines, to service sites, such as loans and book searches in digital libraries.

Sniffers/Wireshark

One way to verify the functioning of computer networks is by means of tools capable of monitoring the flow of data passing through networks at various levels of the OSI model in real time, such tools being called packet analyzers of communication networks or sniffers. These tools run on some networked device that passively receives all data packets from the link layer. After capturing the data that is addressed to the machine, these can be saved for later analysis (Asrodia and Patel, 2012).

Sniffers can be used to convert binary data into a human-readable format, analyze network performance, detect network intrusion, detect spyware, and learn about protocol performance in computer networks (Orebaugh and Ramires, 2004). According to Asrodia and Patel (2012), in addition to the use of sniffers for traffic monitoring and analysis, this use provides several solutions for problems with computer networks. However, they can be a security threat to the individual, because of their ability to capture all incoming and outgoing network traffic, including passwords and usernames or other sensitive data.

In this study, we used Wireshark, free software based on the General Public License (GPL), which captures and analyzes network packets in real time, displaying in detail the data that is circulating in the computer network. Wireshark is primarily used by: network administrators, to troubleshoot computer networks; security engineers, when they need to examine

problems related to network security; developers, who seek to debug protocol implementations; students and other network professionals who use the tool to learn about internal network protocols (Wireshark, 2017). This tool was used during interaction with Brazil's National Digital Library to verify the possible data collected by the digital environment.

Results

The analysis included the identification of digital libraries in South America, based on the collection phase with the Privacy factor of the DLC of the libraries. As a result, nine countries that have National Digital Libraries were found. However, it was not possible to access the websites of the National Digital Libraries of Bolivia and Guiana because they were not found on server, and the National Digital Libraries of Suriname and French Guiana were not found.

Analysis of digital library sites

It can be seen in Table 1 that only the National Digital Library of Brazil's website and the National Digital Library of Colombia's website present some orientation regarding privacy. Most of the libraries use HTTP (HyperText Transfer Protocol), configuring issues with data security and consequently threats to privacy and protection of personal data, except Argentina and Brazil. Three libraries (Argentina, Brazil, and Chile) request some type of registration to reserve documents and, this broadens the set of data about the user and possible implications in the privacy of individuals.

The collected data that are explicit to the users are those requested at the time of registration, authentication to access a service or the search term for retrieval of documents or books.

Data collection using Wireshark

When using the Web, the user requests service based on the client-server model, in which the user requests information and the server responds. Between the lines of this process, the data collection is present, passing through the layers of the OSI reference model. Evidences of privacy threat and levels of abstraction are presented in the next topics in the data collection phase, through access to the National Digital Library of Brazil's website.

To demonstrate the process, the National Digital Library of Brazil page was requested through a query to retrieve a particular book; this activity does not require the user to be logged into the system. In this way, the only data that the user makes available consciously and voluntarily is the search term. The user must log into the site if they wish to reserve books or documents.

This process was accompanied by the Wireshark software and resulted in the collection of 498 packages, of which 267 were directly identified as user interaction packets with the library site. Of this total, 117 packets sent from the originating machine (user) to the server (library page) and 187 packets sent from the server to the user's machine.

To perform this study, package 68 was selected, corresponding a POST method, whose purpose is to allow the user to send data to the server, in this case, to perform the search in the digital collection. The description of the fields obtained during capture with the Wireshark tool follows.

In the Frame field, the metadata of the selected packet relative to capture information, time variables (such as the date and time the packet was captured and the time at which the packet was collected), package size, and protocols are specified acted in this package. In this layer, a GUID (Globally Unique Identifier) is defined in the field "interface id", value generated by the operating system in order to create a unique reference number for the resource (Figure 3).

The Ethernet II field, (Figure 4), is related to the proposal of the data link layer, in order to be the path understood between the origin and the destination, transporting data packets through protocol.

For identification of the source and destination device, the MAC address (Media Access Control), a unique address of the board, is collected. In this case, the MAC address of the Src user card: HonHairPr_f8: b1:51 (xx: xx: xx: x: xx: xx) and the destination MAC address Tp-LinkT_15: e5: 66 (xx: xx: xx: xx: xx: xx).

The Internet Protocol Version 4 field (Figure 5) represents the network layer, through which it selects paths, so that data packets can travel. To do this, it uses the IP (Internet Protocol) address, and in this way, the packets are identified through the source and destination IP address. Geolocation data is also specified for the source and destination, using the Source GeoIP and Destination GeoIP fields.

The Transmission Control Protocol (TCP), Figure 6, refers to the transport layer of the OSI model, in order to allow communication between programs or processes through the port number. Note the presence of the TCP, which carries out the communication through the Src Port: 55498 (55498) and the destination port Dst Port: us-cli (80).

The HTTP is related to the application layer of the OSI model, the only layer typically perceived by the user, which, through the HTTP, allows communication between browsers and servers. Therefore, HTTP

Table. 1. National Digital Library – Countries of South America.

Name	Country	Privacy Policy (Explicit)	Protocol	Collected data (Explicit)	Need of requests registration for download or preview
Bibliotec Nacional Mariano Moreno[1] (Digital Collections)	Argentina	No	https[2]	Contact information (name, email, subject, destination, message) Reservation (login and password) Search term	No However, request login for material reservation
Biblioteca Nacional Digital Brasil [3]	Brazil	No The solely available policy is provided by Disqus service, referring to the data collected in the comments about the document	https	Contact us (Name, email, subject, message) To receive updates (email) Search term	No However, request login for material reservation
Biblioteca Nacional Digital de Chile[4]	Chile	No	http	Contact data (Name, last name, gender, region, occupation, email, action, comment)	No However, it requests registration for the option "request copying" of images with better resolution
Biblioteca Nacional de Colômbia[5]	Colombia	It is presented in the link 'citizen Service' guidelines on personal data	http	Material reservation (email and password) Subscribe citizen service (Type of application, Preferential attention, First name, Second name, Surname, Second surname, Document type, Document number, Address, Optional fixed phone, Cell phone number, Optional cell number, Email, Optional email, country, Department, City, Description of the request), Search term	No
Biblioteca Nacional del Ecuador Eugenio Espejo[6]	Ecuador	No	http	Subscribe to the repository (email and password) Search term	No
Biblioteca Nacional Paraguay[7]	Paraguay	No	http	Search term	No
Biblioteca Nacional del Perú[8]	Peru	No	http	Contact data (name, email, message) Search term	No
Biblioteca Nacional D Uruguay[9]	Uruguay	No	http	Email and password Search term	No
Biblioteca Digital de Venezuela[10]	Venezuela	No	http	Data for Comments (name, subject, comment) Search term	No

1.Available at: https://www.bn.gov.ar/ (accessed 3 July 2017).

2.It is secure HTTP, security principles are added, so clients send confidential information to servers (Belshe and Peon, 2015).

3.Available at: https://bndigital.bn.gov.br/ (accessed 3 July 2017).

4.Available at: http://www.bibliotecanacionaldigital.cl/bnd/612/w3-channel.html (accessed 5 July 2017).

5.Available at: http://catalogoenlinea.bibliotecanacional.gov.co/client/es_ES/bd. (accessed 6 July 2017).

6.Available at: http://repositorio.casadelacultura.gob.ec/ (accessed 6 July 2017).

7.Available at: http://bibliotecanacional.gov.py/bibliotecadigital/ (accessed 10 July 2017).

8.Available at: http://bdigital.bnp.gob.pe/Bvirtual/Home (accessed 10 July 2017).

9.Available at: http://bibliotecadigital.bibna.gub.uy:8080/jspui/ (accessed 10 July 2017).

10.Available at: http://bibliotecadigital.bnv.gob.ve/ (accessed 12 July 2017).

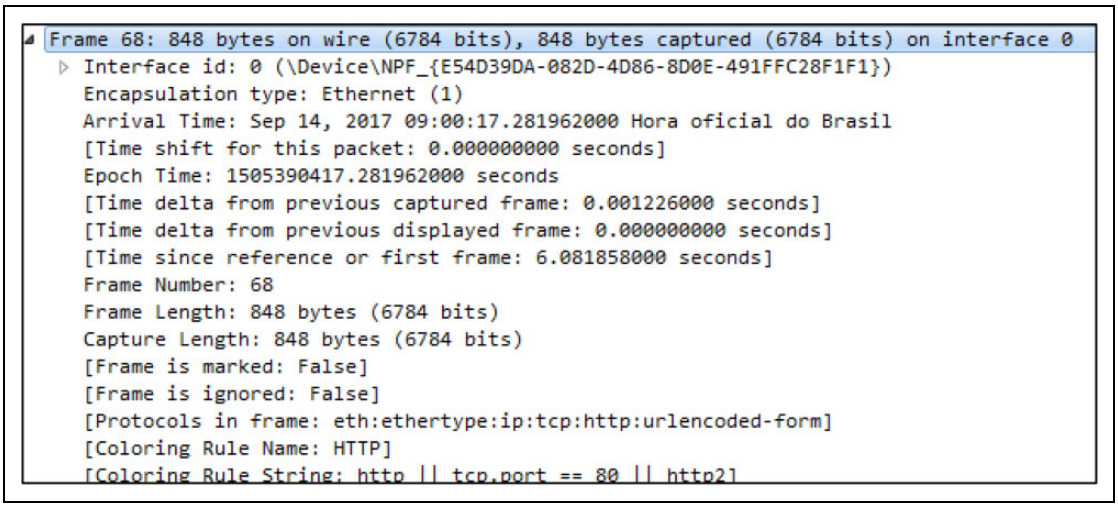

Figure 3. Trimming the Frame field in Wireshark.

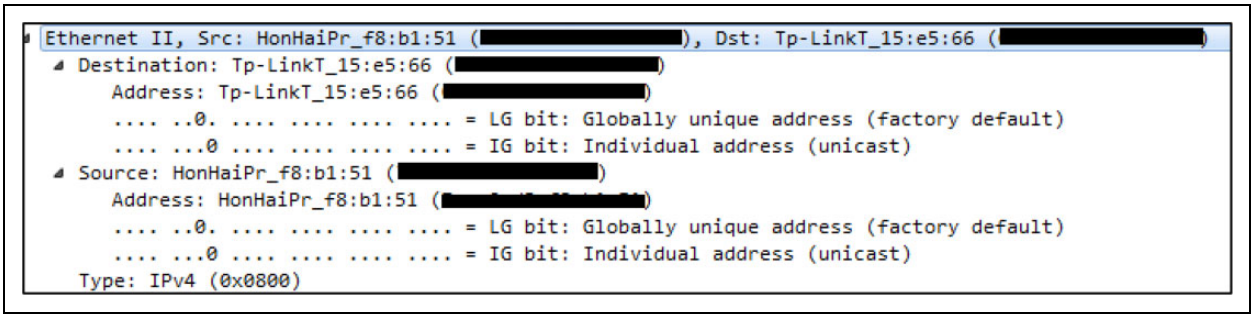

Figure 4. Trimming the Ethernet II field in Wireshark.

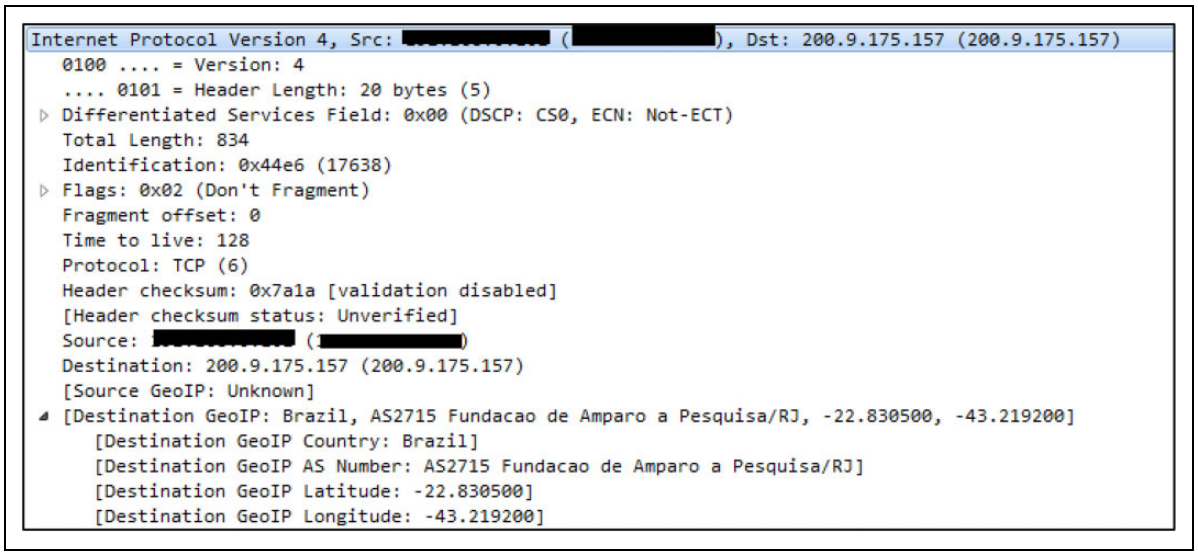

Figure 5. Trimming the Internet Protocol Version 4 field in Wireshark.

is used to send application-layer commands between client and server.

Using the POST command, the client (user) sends a package to the server. This command is used, when the user fills some form in the page (in this case, text to perform the search). Among the information specified in the POST method are: the Uniform Resource Identifier (URI) of the library site, the address to which the data is being sent; user-agent header,[13] with browser and operating system information; referrer,

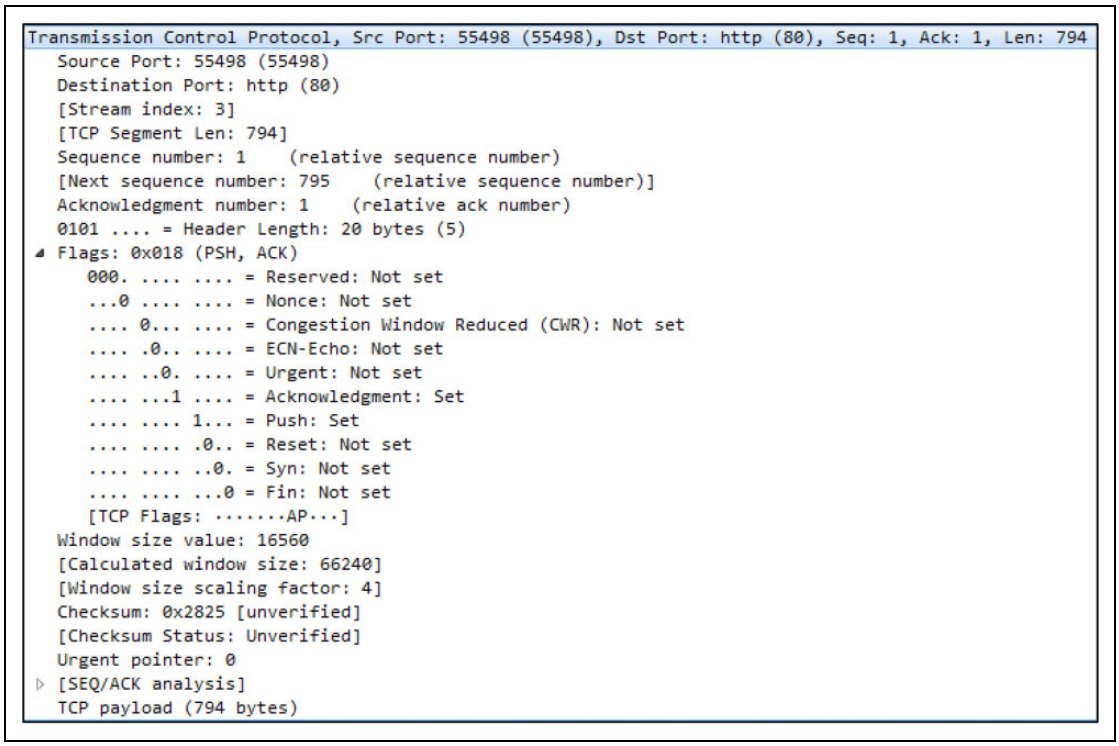

Figure 6. Trimming the Transmission Control Protocol field in Wireshark.

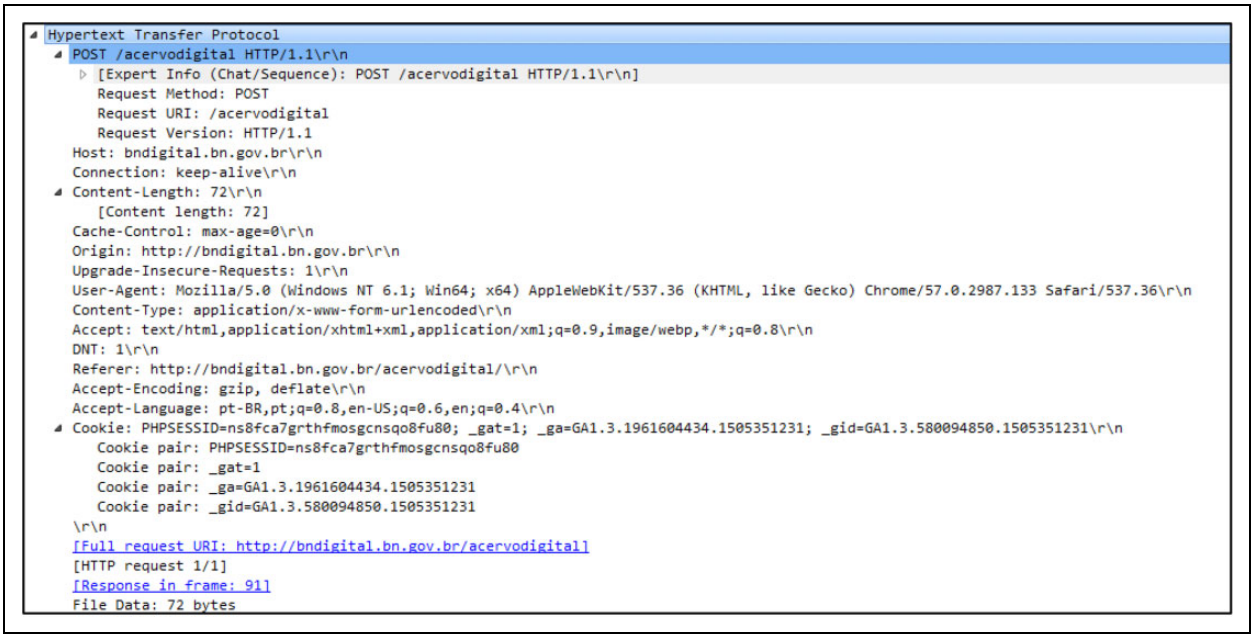

Figure 7. Trimming the HyperText Transfer Protocol field in Wireshark.

which indicates the URL (Uniform Resource Locator) requested, and the accept-language header, which informs the server the language the client machine will be using (Figure 7).

Subsequently, the search terms sent to Brazil's National Digital Library are displayed, explicit in the HTML Form URL Encoded field of the package (Figure 8). It is observed in Figure 8 that the data presented are the ones that the user made available in the search field, data that are in the application layer, the one closest to the user, allowing awareness and apparently do not cause privacy threats when used alone.

Discussions

We analyze possible data collected by the National Digital Libraries sites in two ways: through the

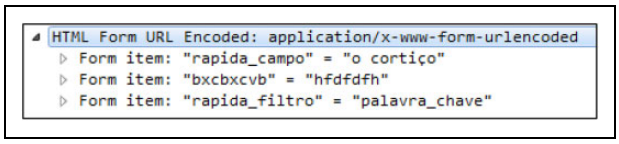

Figure 8. Trimming the HTML Form URL Encoded field in Wireshark.

exploration of the sites and identification of which data can be collected, and analysis of a package of traffic data referring to the user interaction Brazil's National Digital Library website. In the exploration of the sites it was observed that the availability of privacy policies in digital libraries, which are essential to promote the user's awareness about the data collection process, is limited. Additionally, most sites operate under the HTTP which does not provide guarantees regarding the confidentiality and privacy of the data.

Regarding data collection with the use of Wireshark, it is possible to verify data that are present during the user interaction with the digital environment. The data were collected in only one package range from the request date and time, IP address, location data, browser and operating system information, cookies, and machine MAC address and number of ports for communication. These data are not perceptible to the user at the moment of interaction with the environment, confirming the asymmetry of information between the holder and user. The perceptible data are only those that are reported by the user, such as e-mail, registration data, and search term, as shown in Table 1.

Regarding privacy threats in the data collection phase, Table 2 presents a summary of the main data present in the user-server interaction packet when accessing Brazil's National Digital Library site. Each data attribute is classified by its privacy type (Identifiers, Quasi-Identifiers, Sensitive and Not Sensitive).

The MAC address represents a unique and immutable value that allows the identification of the user's machine. The search term and cookies are considered sensitive data, since they store information that refers to something particular to the individual, and that if used improperly can put the subject referenced in these data in situations of embarrassment. However, although IP data, geolocation data, accept-language header, source port, destination port, user-agent are not data that allow uniquely identifying the individual, when correlated with other databases, the examination may result in the identification of the individual.

Through the user-agent header, each time the user interacts with digital environments, this type of data reveals exactly the browser that the user is using and some more data. This information when combined, for example, with location data, can help distinguish users from each other's Internet, making it easier to fingerprint to track on the Web.

Figure 9 illustrates the data collection process and abstraction levels, represented by the layers of the OSI model. This process starts at the time of the client's request (source) to an HTTP page or to an HTTPS page, in which the interaction of the user with the environment depends on the data it provides for the application (conscious interaction process). The architecture of computer networks, divided through the layers of the OSI reference model, determines the interfaces where abstraction is present. This abstraction occurs through the encapsulation of the data collection effected by the protocols that provide the transition of data between the layers, in which is circulated an amount of data that can threaten the privacy of the user, as shown in Table 2. Thus, the very interface of computer networks can contribute to decreasing the user's perception of data collection, making privacy issues more tense and complex.

This research sought to emphasize the possible data collected by digital environments during user interaction. In the case of digital libraries, it was observed that the data collection refers to the data of registers and search term – a situation that is explicit to the user during the interaction with the pages of the digital libraries. However, with the analysis of network packets, it is noted that many other data can be collected and are not perceptible to the user, such as IP address, user-agent header, geolocation, and cookies.

Thus, the user's perception about data collection is based on the data made available. It is not explicit that, encapsulated in the network layers, other data are collected and can threaten privacy, increasing the abstraction for the user about this process. From a different perspective, the data keeper's knowledge of the data across the network layers is increased. However, through this process, other data subjects are intercepted resulting in new or even unwanted collections and emerging privacy threats for individuals.

Most environments do not provide privacy policies, which can contribute to minimizing user insight on the data collection phase. Digital environments should in their content make explicit not only the collection of data that are easily noticeable to the user but also the data that are present in the flow of communication through computer networks.

Considerations

In this study, we have highlighted the privacy issues in the collection phase in digital library sites,

Table. 2. Synthesis of the main data present in the user-server package.

Access with http protocol					
Field	OSI Layer	Atributte	Value	Data	Awareness
Ethernet II	Data link	Source MAC	HonHairPr_f8: b1 (xx: xx: xx: xx: xx)	I	Low
		Destination MAC	Tp-LinkT_15: e5:66 (xx: xx: xx: xx: xx: xx)	I	
Internet Protocol	Network	Sorce IP	IP xxx.xxx.x.xxx	QI	Low
		Destination IP	IP 200.9.175.157	QI	
		Source GeoIP	*Unknown*	QI	
		Destination GeoIP	Brazil, AS2715 Fundação de Amparo à Pesquisa, Latitude: - 22.830500, Longitude: - 43.219200	QI	Low
Transmission Protocol Version	Transport	Source Port	Src Port: 52368(52368)	QI	Low
		Destination Port	Dst Port: us-cli (80)	QI	
Hypertext Transfer Protocol	Application	User-Agent	Mozilla/5.0 (Windows . . . .)	QI	Low
		Accept-Language	Pt-BR\r\n	QI	Low
		Cookies	PHPSESSID . . .	S	Low
		Form item: rápida_campo	o cortiço	S	High
		Form item: rápida filtro	Keywords	NS	

I: identifier; QI: Quasi-Identifier; S: Sensitive; NS: Not Sensitive.

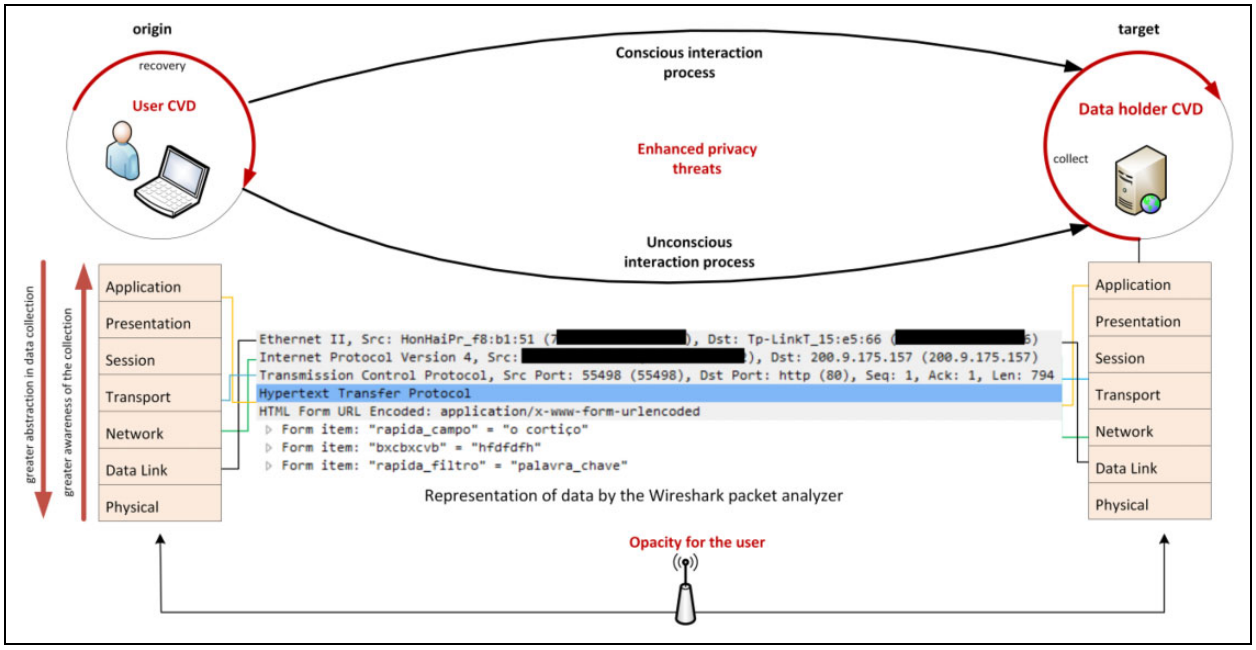

Figure 9. Data collection process.

analyzing the data that are collected both explicitly and implicitly. For this, an exploratory research was carried out in the sites of the National Digital Libraries of South America, and in the analysis from the traffic data resulting from the interaction of the user with Brazil's National Digital Library. By organizing and simplifying their complex context, abstraction layers encapsulate details of communication in computer networks, generating hidden details about collection flows to which users are unknowingly inserted, and increase the privacy-related issues of individuals referenced in sets of data.

Thus, it is observed that the opacity in this scenario goes beyond the low awareness of the user about the collection process and may imply threats in privacy issues since data processed in the networks can result in the identification of the individual. Other aspects can also be of concern, such as the possible correlation of the data with other databases, forming user profiles and increasing the knowledge of the data holders regarding the user.

In conclusion, privacy issues can be influenced by the user's low awareness of when, how and where data collection takes place. Digital libraries need to make privacy policies available for the purpose of guiding users in relation to data collection ensuring that these policies not only specify data that users voluntarily provide, but also data that is abstracted into the layers of computer networks.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Elaine Parra Affonso http://orcid.org/0000-0002-3953-462X
Ricardo César Gonçalves Sant'Ana http://orcid.org/0000-0003-1387-4519

Notes

1. A concept based on the asymmetric information theory developed by Akerlof (1970), which analyzes the implications of asymmetric information in used car markets, in which the seller of a car knows more than the buyer about the quality of that product.
2. Download available at: https://www.wireshark.org/download.html
3. Identifies the user's browser and provides certain operating system details to the servers that host the sites that the user visits (MICROSOFT, 2017).

References

Akerlof G (1970) The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 83(3): 488–500.

ALA (American Library Association) (2017) Privacy. Available at: http://www.ala.org/advocacy/privacy (accessed 15 July 2017).

Asrodia P and Patel H (2012) Network traffic analysis using packet sniffer. *International Journal of Engineering Research and Applications* 2(3): 854–856.

Belshe MPR and Peon R (2015) Hypertext Transfer Protocol Version 2 (HTTP/2), RFC7540. Internet Engineering Task Force (IETF). Available at: https://tools.ietf.org/html/rfc7540 (accessed 3 August 2017).

Facebook (2017) Políticas de privacidade do Facebook. Available at: https://www.facebook.com/privacy/explanation?pnref=lhc (accessed 8 May 2017).

Fabian B, Goertz F and Kunz S et al. (2010) Privately waiting: A usability analysis of the Tor anonymity network. In: *Sustainable e-business management: 16th Americas conference on information systems, AMCIS 2010*, Lima, Peru, 12–15 August 2010, pp. 63–75. Available at: https://www.researchgate.net/publication/220893134_Privately_Waiting_-_A_Usability_Analysis_of_the_Tor_Anonymity_Network (accessed 16 June 2017).

GDPR (General Data Protection Regulation) (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN (accessed 3 July 2017).

Klinefelter A (2016) Reader privacy in digital library collaborations: Signs of commitment, opportunities for improvement. *I/S: A Journal of Law and Policy for the Information Society* 13(1): 199–244. UNC Legal Studies Research Paper. Available at: http://scholarship.law.unc.edu/cgi/viewcontent.cgi?article=1027&context=faculty_publications (accessed 8 July 2017).

Mayer-Schönberger V (2011) *Delete: The Virtue of Forgetting in the Digital Age*. Princeton, NJ: Princeton University Press.

Microsoft (2017) User agent. Available at: https://msdn.microsoft.com/en-us/library/hh920767(v=vs.85).aspx (accessed 3 June 2017).

Nikkel BJ (2005) Generalizing sources of live network evidence. *Digital Investigation* 2(3): 193–200.

Nissenbaum H (2011) A contextual approach to privacy online. *Daedalus* 140(4): 32–48.

O'Hara K and Shadbolt N (2014) *The Spy in the Coffee Machine: The End of Privacy as We Know It*. London: Oneworld.

Orebaugh AD and Ramirez G (2004) *Ethereal Packet Sniffing*. Syngress.

Samarati P (2001) Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering* 13(6): 1010–1027.

Sant'Ana RCG (2013) data life cycle for Information Science (DLC-IS). In: *Encontro Nacional De Pesquisa Em Ciência Da Informação*, 14, 2013, Florianópolis. Anais . . . Florianópolis. Available at: http://enancib.sites.ufsc.br/index.php/enancib2013/XIVenancib/paper/viewFile/284/319 (accessed 3 January 2017).

Sant'Ana RCG (2016) Data life cycle: A perspective from the Information Science. *Information & Information* 21(2): 116–142.

Sclavos J, Simoni N and Znaty S (1994) Information model: From abstraction to application. In: *IEEE network operations and management symposium*, Orlando, Florida, USA, 14–17 February 1994, p. 183. IEEE.

Spiekermann S and Cranor L F (2009) Engineering privacy. *IEEE Transactions on Software Engineering* 35(1): 67–82.

Tanenbaum AS (2003) *Computer Networks*. 4th edn. [translated edition]. Rio de Janeiro: Pearson.

Tanenbaum A S and Wetherall J D (2011) *Computer Networks*. 5th edn. Rio de Janeiro: Pearson.

Tavani H T (2008) Informational privacy: Concepts, theories, and controversies. In: Himma KE and Tavani HT (eds) *The Handbook of Information and Computer Ethics*. Hoboken, NJ: John Wiley & Sons, pp. 131–164.

Wireshark (2017) User Manual. Available at: https://www.wireshark.org/docs/ (accessed 8 July 2017).

Word Digital Library (2017) Warnings from the World Digital Library. Available at: https://www.wdl.org/pt/legal (accessed 8 September 2017).

Author biographies

Elaine Parra Affonso is a doctorate student in Information Science at São Paulo State University (UNESP), School of Philosophy and Sciences, Marília, São Paulo State, Brazil. She has a Master's in Computer Science and is Professor in the Technology School of Presidente Prudente (FATEC), São Paulo State, Brazil.

Ricardo César Gonçalves Sant'Ana is Doctor in Information Science at São Paulo State University (UNESP). He is Adjunct Professor at UNESP, Campus Tupã and Professor in the Post-Graduate Program in Information Science at UNESP School of Philosophy and Sciences, Marília, Brazil.