Author's response to reviews

Title: A RESTful interface to pseudonymization services in modern web applications

Authors:

Martin Lablans (lablans@uni-mainz.de)
Andreas Borg (borga@uni-mainz.de)
Frank Ückert (ueckert@uni-mainz.de)

Version: 4 Date: 20 November 2014

Author's response to reviews: see over
Dear Ms. Pura,

attached you find a revised revision of our manuscript “A RESTful pseudonymization interface for use in modern web applications”. Following a suggestion of reviewer Muñoz Carrero, we changed the title to “A RESTful interface to pseudonymization services in modern web applications”.

In the following, we state our responses to the reviewers’ comments. We also uploaded a PDF file of the manuscript where the changes of this revision are marked.

Best regards

(Martin Lablans)  (Andreas Borg)
Reviewer Adolfo Muñoz Carrero:

Change the title to clearly reflect the content of the paper. Something like that could do the trick: A RESTful interface to access pseudonymization services for modern web applications.

We changed the title to “A RESTful interface to pseudonymization services in modern web applications”

Clarify the objective in the background section. In page 2, line 2 it is said that one of the problems to fix in order to be able to use the information for research is the pseudonymization of the records. This is not solved by simply changing the ID of the patient: The ID or the name can be found in several parts in the record, like in the multimedia information, in free-text sections or in information regarding one of the patient’s relatives. The identity can even be established combining some other data (quasi-identifiers) like birthdate, sex, and home place. So, in order to keep privacy, the record must be filtered and modified before it can be safely disclosed. Probably, the objective of this paper is not to solve this question; however, it should clarify that to avoid misunderstandings in the audience.

In the introduction, we now make clear that this is out of focus of this paper:

“It must be clarified that our solution is limited to pseudonymization in the sense of replacing IDAT in the form of well-defined attributes. Related tasks like removing embedded IDAT from image data or free text forms, or preventing re-identification of patients through quasi-identifying attributes in the MDAT are research areas of their own and out of focus of this article.”

Transference of sensible information could be dangerous because of the risk of interception. In background section it could be necessary to specify if there is some safeguard implemented to protect the communication against that risk.

As the interface uses HTTP, the same security considerations as for internet connections apply. We added the following remark in section 2.1 (“RESTful web services”):

“Also, network connections can easily be protected using Transport Layer Security (SSL/TLS) to secure transmission of sensitive data via the proven HTTPS protocol. Separating transport security from the interface allows administrators to freely choose reasonable security measures to fulfil the requirements of a specific application, ranging from encryption to more complex solutions like virtual private networks.”

After reading the manuscript it seems that all demographic data are in the IDAT while all medical data reside in the MDAT, which is the only information accessible for secondary use. This could not be enough for a majority of research usages where knowledge about age, sex or residence of patients is necessary in order to extract valid conclusions. On the other hand, it is true that these data can lead to the re-identification of the patient. This risk is treated through several models (k-anonymity, l-diversity, t-closeness) and there are techniques (generalization, removal of unusual values ...) to
minimize it. We think that the paper should deal with this subject or the authors should justify the elimination of all the demographic data.

Concerning the risk of re-identification, see the above addition to the introduction. Regarding the use of demographic data in the MDAT for research, we now clarify:

“Also, our distinction between IDAT and MDAT does not preclude that the MDAT may contain some amount of demographic data (for example sex or age), provided that this does not significantly increase the risk of re-identification.”

Figure 2 shows the components of the Mainzelliste reference implementation. The relationships depicted inside the Patient List components lack of clarity and might cause ambiguity. We think that this figure should be improved to show in a clearer way how these components interact.

We reorganized Figure 2 in a way that should make the interaction of components more clear. Also, in order to provide a better link to the explanatory text, we added to the figure label: “See section 3.2 for a description of the components and their relations.”

There is an ISO standard, the Technical Specification ISO/TS 25237 that defines the concepts for identification and pseudonymization (along many other things). It could be interesting to check it and to state in the paper that whenever the authors use the pseudonymization related concepts in the text their meanings are those defined by this standard.

We aim at free distribution and use of our interface and the reference implementation by publishing them under “free” licenses (Creative Commons and AGPL) and deliberately submitting this article to an open access journal. We believe that referring to a closed standard such as the ISO specification in such a fundamental way would counteract these aims. Nevertheless we are thankful for the tip and will double check if we missed knowledge there.
Reviewer Wolfgang Hoffmann:

P 4, line 28-30 “... up to two errors can be detected ...” “...the transposition of two adjacent characters can be corrected...” how is this achieved? how can be excluded that the respective pseudonyms with these kinds of differences do not in fact refer to different patients?

Such detection and correction of errors is a feature of the underlying mathematical code. We now added “Faldum and Pommerening further show that...” to the statement to make clear that we rely on their work.

P 4, l 37 what is meant with “... in an error tolerant way...” obviously, the degree of deviance, that can be handled by any algorithm is limited. How are these limits defined? is there empirical evidence to define equality vs. inequality of the identifiers in individual records?

This refers to the fact that record linkage should deal with spelling errors and similar errors as shown in the example. To avoid confusion, we removed the phrase “in an error-tolerant way”. We further added two references to comprehensive works on record linkage for readers who want to learn more about the subject. A broader discussion on the topic would be out of focus of this article.

P 4, l 45-46 I do not understand, why an additional “ability” renders a pseudonymization service less simple than one without this ability – do the authors rather mean “demand”?

To clarify this point, we changed this passage to:

“However, the ability to identify patients across multiple data sources is an essential motivation to delegate pseudonymization to a centralized service instead of generating pseudonyms locally. We therefore recommend that every implementation of the interface provide a reasonable record linkage algorithm.”

P5, l 31-33 How can this advisable feature be implemented in the typical use cases? Shouldn’t it be enforced?

The Mainzelliste interface or its implementations cannot enforce how sessions are mapped to users as it is user-agnostic by design (it only “knows” other servers, not individual users). We added a clarification and a suggestion concerning the implementation on the MDAT server:

“It is advisable that the MDAT server mirror every browser session with a session on the patient list server, although this cannot be enforced due to the fact that authentication of individual users is a duty of the MDAT server (see bullet point “Possibility to delegate user authentication” in section 1.1). Mapping sessions to users can easily be implemented by creating a session once a user logs in and storing the session id in the user account data”

P9, l 21-37 The text reviews existing solutions from a single and current technical perspective. As different projects have different requirements on architectures and design, a more balanced review seems recommendable. E-PIX, for example, was designed a couple of years ago as part of a complex
hospital-internal research platform, with no bandwidth or latency bottlenecks to consider, based on Java EE and ESB. Later platform and E-PIX extensions have added capabilities for the RESTful interface leaving the tool itself as-is

We added two sentences that stress E-PIX' origins in a clinical application and the different requirements that arise therefrom: “However, one must keep in mind that E-PIX was initially developed for use in an application where integration into the existing clinical IT infrastructure is important and no significant network limitations (bandwidth, latency) had to be considered.” (Concerning the use of SOAP), “Once again, the different approach of E-PIX is motivated by its use in a clinical environment and seems reasonable for the originally intended clinical use case.” (Concerning the authentication of users in the IDAT system).

P8, l 4 Programming of a new tool inspired by the shortcomings of an existing one does not instantly make such new tool the “successor” of the existing one. Shortcomings of the existing one can easily inspire a multitude of new tools being designed and implemented. To (formally) become the successor of the existing one requires more than added features and running code. The statement made should consider strategic and technical aspects of the cooperative development of the PID generator 2.0 in the TMF community, including the option of not one tool being the 1:1 successor.

Mainzelliste succeeds the PID Generator's as a pseudonymization tool in our own research projects. Of course we cannot claim it being a successor in general. We changed the sentence to “The PID generator and its drawbacks were a major motivation for developing Mainzelliste as a possible replacement.”

P 3, l 43 please spell out UUID at first mention

P 10, l 34 Switzerland

These issues have been corrected.

P 1, l 4 (and further on) The article does not clearly distinguish between identity management with the assignment of some identifier as representation for the identity and pseudonymization referring to a domain-specific assignment of another identifier (the pseudonym) to protect the identity of a person. The discussion of the RESTful interface mixes aspects of both throughout the discussion. An explicit definition would improve the consistency of the arguments.

We clarified this by stating in the introduction:

“This type of pseudonymization, which we call “first-level pseudonymization”, is to be distinguished from the case where an existing pseudonym is transformed into another, for example when patient identifiers in different domains need to be translated into one another (see the discussion of PIX/PDQ in section 4.1.2). In this paper, we will refer to first-level pseudonymization if not stated otherwise.”

Also, the ability of the interface to allow second-level pseudonymization is made clear at the end of section 3.1.2:
Another use case that can be implemented by this token type is the transformation of one pseudonym to another ("second-level pseudonymization"). For this purpose, the token can be configured to allow reading pseudonyms of a specified type (i.e. domain).

P 10, l 25 The German National Cohort operates an identity management that is, in fact, not based on the Mainzelliste but on E-PIX. However, the National Cohort’s system uses a RESTful interface which is compatible with the Mainzelliste RESTful interface to a large extent.

This listing refers to implementations of the interface, not to use of the reference implementation. In order to make that more clear we changed the sentence to “The described RESTful pseudonymization interface has, in addition to the reference implementation, already been implemented by…” and restricted the statement concerning the National Cohort insofar as it implements the interface “to a large extent.”