

Pseudonymous Identifiers and the Data Protection Directive (95/46/EC)

Document Version: 05 DRAFT

Date: Dec 2008

Author: Andrew Cormack (JANET(UK))

Abstract

This paper suggests measures by which the status of pseudonymous identifiers – whether they are personal data or not – under the European Personal Data Directive (95/46/EC) may be able to be controlled by the issuers and users of those identifiers.

The paper does not constitute legal advice, and no responsibility is accepted for any errors. Readers should note in particular that the case law and opinions in this area are particularly unclear, and may vary between different European countries.

Pseudonymous Identifiers

The European Data Protection Directive (95/46/EC) defines “personal data” in Article 2 as follows:

“(a) 'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;”

However there is uncertainty over whether pseudonymous identifiers, which conceal users' identity but allow them to be recognised on a return visit, are inherently personal data or not. This is a particular issue for the unique “handles” issued by some federated access management systems to allow a service to save a user's search and preference information between visits. These are designed to preserve privacy as far as possible, for example the definition of the widely used eduPersonTargetedID pseudonymous identifier states that:

“This attribute is designed to preserve the principal's privacy and inhibit the ability of multiple unrelated services from correlating principal activity by comparing values. It is therefore REQUIRED to be opaque, having no particular relationship to the principal's other identifiers, such as a username or eduPersonPrincipalName. It SHOULD be considerably difficult for an observer to guess the value that would be returned to a given service provider.

It MAY be a pseudorandom value generated and stored by the identity provider, or MAY be derived from some function over the service provider's identity and other principal-specific input(s), such as a serial number or UUID assigned by the identity provider.”

These identifiers therefore permit a service to recognise a series of visits by the same user, but should not allow the service to identify the real-world individual that made those visits.

Personal Data

A different class of identifiers having similar characteristics, IP addresses, was considered in the Article 29 Working Party's Opinion 4/2007 on the Concept of Personal Data which considered that:

“unless the Internet Service Provider is in a position to distinguish with absolute certainty that the data correspond to users that cannot be identified, it will have to treat all IP information as personal data, to be on the safe side.” (page 17)

Much of the discussion of that statement has concentrated on the circumstances that might make it technically certain that an IP address could not be linked to an identified person. Typically this will only occur if the network access provider does not know (or does not record) the identity of those to whom it issues IP addresses, for example an open wireless

access point that accepts connections from any passing laptop or a cybercafé that does not require individual users to identify themselves.

However it has now become apparent that legal, as well as technical, measures could be used to prevent linkage, and that this might satisfy the requirements of the Directive to make the identifiers not personal data. This approach has been taken by the EC-funded ACGT medical research project where hospitals provide patient data under contract to an independent third party. The third party replaces all identifiers with pseudonymous ones, in a way that permits researchers to link together records from the same patient but ensures that only the third party can identify the real-world individual to whom the linked records refer. This is technically similar to many pseudonymous identifier systems. However the third party in this case also has a legal agreement with the project board (which acts as data controller), stating that it will not disclose the linking information to any researcher. This is considered sufficient to permit the researchers to treat their data, associated only with an unlinkable pseudonymous identifier, as non-personal data. The Article 29 Working Party Opinion confirms that in circumstances where:

“re-identification of the data subject may have been excluded in the design of protocols and procedure ... processing may thus not be subject to the provisions of the Directive.” (page 20)

This approach, of ensuring by technology that only one organisation can make the link to a living individual, and then ensuring by agreement that that organisation will not do so, also appears to be applicable to federated access management systems. Indeed the Opinion specifically mentions irreversible cryptographic hashes, which are often used to calculate eduPersonTargetedID, as an example of an “appropriate technical measure” that may be used in such cases.

Pseudonymous Identifiers in Federated Access Management

In federated access management systems a user is associated with an organisation known as their Identity Provider (in education this is normally their school, college or university) which verifies their identity using an authentication technology such as a username and password. The Identity Provider then assigns pseudonymous identifiers to each user that Service Providers can use to recognise, but not identify, individual users. As noted above the eduPersonTargetedID specification includes a further rule, that different values of the identifier should be used for different service providers, to prevent them aggregating information provided to different services. Thus the technical measures ensure that only the Identity Provider organisation is able to link a pseudonymous identifier to the real-world individual. It therefore appears that it may be sufficient to ensure that these identifiers can be treated by others as non-personal data for each Identity Provider to have a legal agreement that it will not disclose that linkage. In fact many federations already contain this rule within their membership contracts, for example, participants in the Finish HAKA federation already contract with their federation operator that:

“2.3.9. The Home Organization shall collect a log that includes at least the Shibboleth handle [NameID] and a piece of information that uniquely identifies the End User.

2.4.5. Considering the articles in the Personal Data Act, the Service Provider shall collect a log that includes at least the Shibboleth handle [NameID] of the End User.

To facilitate abuse investigation, the Service Provider shall provide relevant log entries to the Home Organization.”

The practice of Identity Providers in that federation is to investigate and deal with misuse themselves, rather than ever revealing the identity of users to Service Providers other than at the order of a court.

Conclusion and Recommendations

It therefore appears that it may be possible to ensure by legal agreement that pseudonymous identifiers can be classed as non-personal data. Clearly if a service provider subsequently collects information that allows them to link the identifier to the real-world person (for example by asking the user for their name or e-mail address) then the identifier will become personal data, subject to all the compliance requirements of EU and national laws.

Identity Providers and Service Providers should therefore use pseudonymous identifiers according to the following guidelines. These are, in any case, good practice for using such identifiers to protect the privacy of users. They may also be sufficient to ensure that pseudonymous identifiers constitute non-personal data, thereby significantly reducing the regulatory burden on Identity Providers, Service Providers and Users.

- Identity Providers should
 - Construct pseudonymous identifier values in ways that conceal as far as possible the identity of the user, for example by using one-way hash functions and providing different values to each service provider;
 - Declare that they will not disclose the identity of the person to which a particular identifier value was assigned, other than when required by law to do so.
 - In particular, reports of misuse or other problems should be investigated by the Identity Provider, who is anyway most likely to be able to hold the user to account, and not the Service Provider.
- Service Providers should
 - Not collect personally identifying information from a user who was otherwise only identified by a pseudonymous identifier;
 - Not seek to obtain information linking a pseudonymous identifier to a user from any other source; in particular they should not aggregate information collected from different services;
 - Provide evidence to Identity Providers to permit them to investigate and deal with any misuse or other problem in the use of the service.

References

Directive 95/46/EC

<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML>

eduPerson specification

<http://www.nmi-edit.org/eduPerson/internet2-mace-dir-eduperson-200712.html>

Article 29 Working Party Opinion 4/2007 on the concept of personal data

http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2007/wp136_en.pdf

ACGT project and Centre for Data Protection on satisfying regulatory requirements for anonymity in medical research

<http://eu-acgt.org/> and <http://www.privacypeople.org/>

HAKA federation service agreement

<http://www.csc.fi/english/institutions/haka/join/HAKAagr04042005.pdf>