Towards Collecting and Linking Personal Information for Complete Personal Online Identity Modelling

Frans F. Blauw(✉) and Sebastiaan H. von Solms

Academy of Computer Science and Software Engineering,
University of Johannesburg, Johannesburg, South Africa
{fblauw,basievs}@uj.ac.za

Abstract. Online identities of users are fragmented amongst multiple websites and online applications (service providers). These identity fragments will contain critical personally identifiable information. Very often, a user is unaware that their personal information is stored by a service provider. In this paper we describe a model – Personal Information Collection and Collation (PICoCo) – used to assemble identity fragments of users and form a complete identity model of a natural person. The ultimate goal is to allow service providers to verify incoming data, but also for a user to discover where their data is being stored. The description of PICoCo includes collection, classification, collation, validation, verification, and discovery.

Keywords: Service design · Information visualization
Personally identifiable information · Privacy

1 Introduction

As users complete their day-to-day tasks in an online environment, their online identities are spread amongst multiple websites and online applications (service providers). These identities are formed using their personal information that is directly submitted to the service as well as the personal information that is created and shaped by the general usage of the service [11]. Across multiple services, a fragmented model of the user's identity is formed by each service, consisting of multiple identity fragments. This identity fragment will contain critical Personally Identifiable Information (PII) that could potentially be used for legitimate and illegitimate uses, such as rendering a service or identity theft, respectively.

Furthermore, not only can parts of this identity model be outdated as the user fails to use the service or update their personal information, but can be often be forgotten about by the user. That is, a user is unaware that their personal information is still stored by a service provider.

M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 507–522, 2018.
https://doi.org/10.1007/978-3-319-91238-7_40

1.1 Collection of an Identity

In this paper, we will define these parts of Personally Identifiably Information (PII) as an "identity fragment". Multiple identity fragments combined together form an "identity model". This is discussed further in Sect. 2.3. For a complete insight into a person's online identity, multiple parts of personal information will be required.

If these identity fragments were to be collated from multiple service providers, a more cohesive identity model can be formed. This identity model will contain the newest personal information about the individual. Constructing this more comprehensive and broader identity model carries with it a multitude of advantages for both the user as well as the service providers [9]. This collection process is discussed in Sect. 3.1.

As identity fragments are collected, multiple sources can be queried to ensure that the personal information contained within the broad model is the newest, up-to-date information. By querying multiple sources, the user will be able to keep track of all service providers that keep personally identifiable information about them – discussed in Sect. 3.5. As such, a user can make informed decisions about their personal information retained by service providers. Fraud can also be more easily detected, prevented, and mitigated by having a more cohesive, complete identity model by completing verification checks against personal information from multiple service providers. This verification of data is discussed in Sect. 3.6.

1.2 PICoCo Overview

This paper investigates a model for collecting personal information fragments from across multiple service providers. This model will be called the Personal Information Collection and Collation (PICoCo) Model and discussed in Sect. 3.

PICoCo collects raw personal information from the service provider and initiates a classification and identification process. Identity fragments are formed for each user of the service provider. These information fragments are then collated with the existing known personal information of multiple users. New fragments are compared with existing information to determine its newness and accuracy. Updates are brought to the existing personal information and a new, updated, identity model is assembled. This process continues with multiple service providers to construct the most comprehensive identity model possible, for each user.

Classification of personal information considers the sensitivity of information across multiple factors. This classification will affect the resilience as well as the accuracy rating of personal information fragments collected.

Furthermore, PICoCo allows for the validation and verification of personal information. Service providers can submit new personal information received from individuals for review. Information received will be validated against the existing identity model constructed for the individual. If the personal information

can be verified to a set level of confidence, the information can then be certified. Otherwise, a potential notice can be returned, and the affected individual can be warned.

This paper discusses further advantages and disadvantages of this model, starting in Sect. 4. One pertinent difficulty is that of linking and collating individual information fragments to bigger models with enough certainty and confidence. Automating the process is the ideal, however, there are times where user input will be required to verify the confidence of potential information linking.

Lastly, further work built on top of this model is briefly discussed in this paper. This includes abstract descriptions of privacy metrics, location-based, and behavioural-based security enhancements.

2 Identity

A person's identity can be defined in a multitude of ways. In this section we will discuss what it is to be a Natural Person (Sect. 2.1) and how a natural person's identity is described in terms of the law (Sect. 2.2).

We will also investigate how we can turn these identities into online facsimiles by describing identity models and the identity fragments (Sect. 2.3 and on) within them.

2.1 Natural Persons

A natural person, in terms of the law, is a living, breathing individual human being [10]. This opposed to legal entities that may be organisations. Each natural person has his or her own, personal, unique identity. In the social sciences, an identity is made up of a person's qualities, beliefs, expression and personality. However, in the information sense, an identity consists of Personally Identifiable Information (PII). That is, information that can be used to identify a specific person to a high confidence. PII is exceptionally prevalent in the information technology sector as the collection thereof can be used to easily track and trace an individual in the online sphere [6].

Even though there is no worldwide standard for PII, there are multiple frameworks that specify precisely what information can identify a specific person [8].

2.2 Natural Persons and Identity in Regulation

The European Union, in 1995, issued Directive 95/46/EC entitled: "Directive on the protection of individuals with regard to the processing of personal data and on the free movement of such data" [4]. This directive not only considers unique identification numbers when referring to personal information, but also includes other factors such as "physical, physiological, mental, economic, cultural or social identity".

An upcoming regulation (Regulation (EU) 2016/679) is set to supersede Directive 95/46/EC in May 2018 and is entitled the "Regulation on the protection of natural persons with regard to the processing of personal data and on

the free movement of such data" [5]. This regulation defines a "data subject" as a digital identity "who can be identified, directly or indirectly, by means reasonably likely to be used by the controller or by any other natural or legal person." That is, digital data, forming an identity, that can be connected to a natural person. Examples given [3] are:

- a name and surname
- a home address
- an email address such as "name.surname@company.com"
- an identification card number
- location data (for example the location data function on a mobile phone)
- an Internet Protocol (IP) address
- a cookie ID (where specific sectoral legislation exists for cookies)
- the advertising identifier of your phone
- data held by a hospital or doctor, which could be a symbol that uniquely identifies a person

The EU also does not recognise the following a personal data:

- a company registration number
- an email address such as "info@company.com"
- anonymised data

Similarly, in the United States of America, the National Institute of Standards and Technology (NIST) have a clear definition of what constitutes PII. NIST published a document in 2010 entitled: "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)" [7]. In Sect. 2.2 of this document, NIST gives a list of clear examples of PII:

- Name (full name, maiden name, mother's maiden name, or alias)
- Personal identification number (passport number, taxpayer number, patient information number, financial account number)
- Address Information (physical, or digital)
- Asset Information (MAC address, IP address)
- Telephone Numbers (personal, mobile, and business)
- Personal Characteristics (photographic, x-ray, fingerprint, and template data for biometric identification)
- Property Identification (vehicle registration)
- Information linkable to the above (date and place of birth, race, religion, weight, activities, geographical indicators)

As can be seen in the publications and standards above, personally identifiable information does not merely refer to identifiers and static facts (name, address, date of birth, etc.) of a natural person. PII also refers to what a natural person owns as well as their activities. Essentially, anything, online and offline, that can be linked to a natural person to some extent, is seen as personal information.

Considering these publications and standards, we can start to form a concept what of constitutes an identity when constructed from personally identifiable information. Each of these pieces of personal information is an element of this identity – an identity fragment.

2.3 Identity Fragments

We will define an identity fragment as a small part of a natural person's personal information that could personally identify him or her. Each fragment on its own will probably not identify a single natural person uniquely to a very high level of confidence. However, combining more identity fragments of the same, unique identity will increase the confidence that a fragment belongs to a single natural person. Many identity fragments added together will form an identity model of a person. This identity model is directly connected to a natural person.

Identity Fragment Collection. Identity fragments can be collected by either active or passive means. The first method is active data collection. This is done by the user supplying their PII to the information system. For example, upon registration, users are often requested to supply information about him or herself. This information is then captured and stored in the information system until such time that the user updates the information, or the information is deleted.

Another means of collecting identity fragments is more passive. As the user uses an information system, data is collected and stored by the information system. These usage patterns can be linked to a natural person and is treated as personally identifiable information (as seen in the previous section). Very often, usage tracking is not limited to a single site, but can be spread across multiple sites using the same information system. Social media integration and advertising networks are a great example of users being tracked across multiple systems and platforms.

Not only are "facts" considered identity fragments, but usage data – data generated as the user actively uses the information system – also forms part of identity fragments. A natural person can merely exist, and patterns and behaviour can then also generate identity fragments. Extreme forms of passive identity capturing exists in elements such as security cameras. With the increase in human recognition technology (such as facial recognition), as a person walks from one point to another, the security cameras along his or her path can see and record this. This visual data can then be used to extrapolate patterns and behaviour of the human subject, again forming an identity fragment.

Completing an Identity Model. Combining data captured from these active and passive means can form a very solid identity model of a natural person. However, normally, not all types of identity fragments exist on a single system. Different identity fragments can be spread across multiple online systems, depending on the use of the system. As shown in Fig. 1, different service providers will have these different identity fragments. Together they can form an identity model of a natural person.

Normally, identity fragments should only be collected if it is required by law and to increase the user's experience using the system. However, it has been seen to not always be the case with data being captured and stored for no apparent reason.

Users normally have no control over data collected via passive means, however users actively providing data have full control over the data being provided. To this end, users can potentially provide fake or misleading data. The verification of data collected via active means is of vital importance to ensure that data can be used effectively and legally.

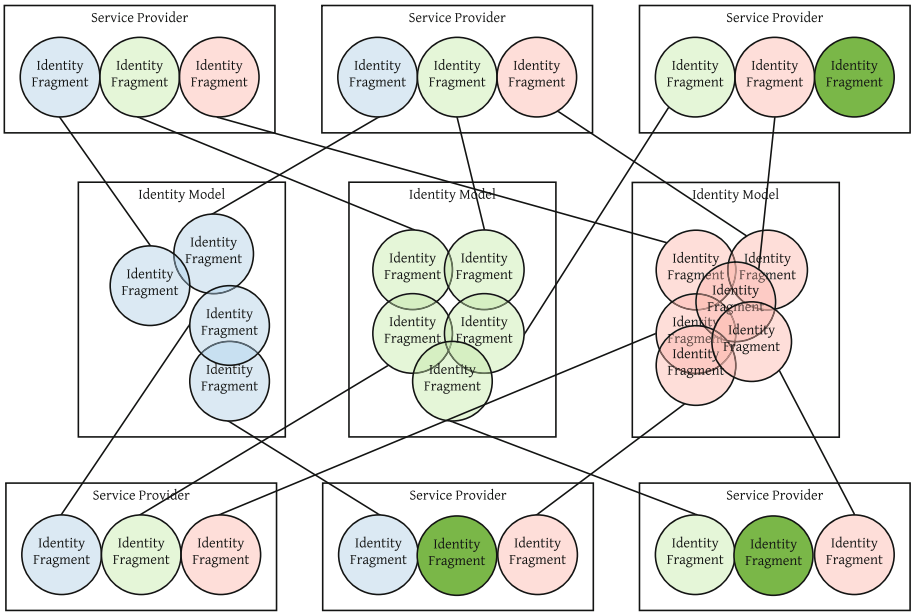

Fig. 1. Identity fragments forming identity models

Stale and Outdated Identity Fragments. Identity fragments collected via active means also have a possibility to become stale, where the data is outdated or irrelevant [2]. If no active updating or verification of data takes place, the keeper of the data has no means of ensuring that data contained in identity fragments are current and still relevant. Certain identity fragments have a longer permanency than others. For example, the likelihood of a person's name or race changing is far lower than his or her address or credit card details.

Identity fragments can also become inconsistent across multiple systems depending on whether the user updates their data when it changes or is prompted to do so. Verification across systems can become difficult when fragments that supposedly belong to a single person is not the same. At the same time, it is difficult to ascertain which data is the most recent and still relevant.

Passive identity fragments can also easily become stale and irrelevant. A slight change in a user's behaviour can immediately invalidate data. New passive collection should validate older data to see whether data is consistent and update it when it needs updating.

However, it should be mentioned that even outdated identity fragments are still considered personally identifiable information. As defined by the standards and regulations in the previous section (Sect. 2.2), historic information of a user is considered PII. Historic information is especially useful to form predictions and perhaps show trends, especially when coupled with more identity fragments.

3 Personal Information Collection and Collation (PICoCo) Model

In this section we will introduce our Personal Information Collection and Collation – or PICoCo – Model. PICoCo consists of six components, each of which has its own function. These components are:

- Collection (Sect. 3.1),
- Classification (Sect. 3.2),
- Collation (Sect. 3.3),
- Validation (Sect. 3.4),
- Discovery (Sect. 3.5), and
- Verification (Sect. 3.6).

Figure 2 gives an overview of the different components within PICoCo and the interaction of different components. Each of these components will be discussed in their relevant sections in terms of their features and how they interact with one another.

This description of PICoCo assumes that all relevant permissions have been obtained and the collection of the identity fragments is completely legal.

3.1 Component 1: Collection

The first part of PICoCo is the collection of identity fragments. As stated in Sect. 2.3, multiple identity fragments make up one larger identity model.

The act of collection will involve a Collector retrieving identity fragments from multiple service providers. This process is shown in Figs. 3 and 4.

First the collector will take an existing identity fragment from a known identity model (steps 1 & 2). That is, an identity model of a natural person that it already knows. Using this known identity fragment, the collector will query each linked service provider (step 3). The service provider queried will look up the identity fragment provided by the Collector and see whether it exists in its database (step 4).

If some data in the service provider's database matches that of the identity fragment provided, the matching data will be highlighted (step 4). Depending on the type of data in the identity fragment, a level of confidence is assigned to this match. The confidence will determine how confident the collector is that the identity fragment provided from the known identity model matches that of the data provided by the service provider (step 5b).

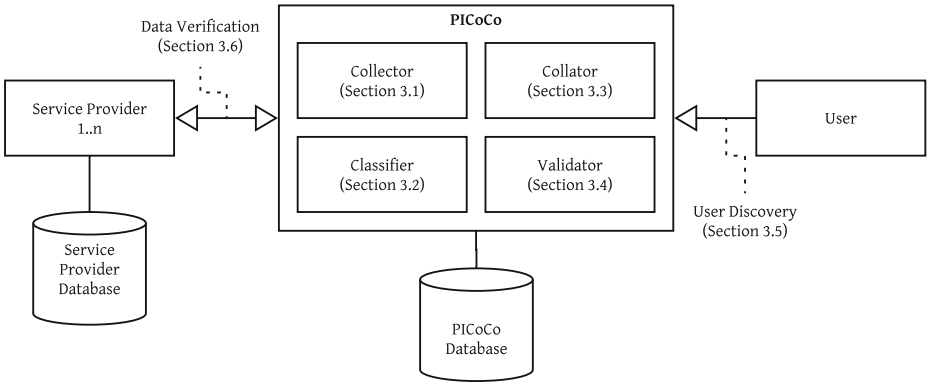

Fig. 2. Personal Information Collection and Collation (PICoCo) Model

If the confidence does not meet a predetermined threshold, more identity fragments will be queried as part of the single identity query (step 6a). If nothing is found, or the confidence remains low, the collection process stops (step 5a).

The collector will query more identity fragments to have the confidence raised above this determined threshold (step 6b).

Once the confidence is satisfactory, the collector will add the queried service provider to the known identity model as a source of data. This will reduce lookup times for future queries, but also allow the user to see which service provider has PII about him or her.

As the collector and service provider now have a matching identity, the service provider will retrieve additional identity fragments that it is linked to the identity being queried (step 7). These additional fragments – unknown to the collector – are provided to the collector (step 8).

The collector will then take matching identity fragments and collate them to form a single identity model of an individual.

3.2 Component 2: Classification

Each identity fragment will carry a weight and signifies the unlikeliness that a particular identity fragment will belong to more than one natural person. This weight will contribute to the confidence of an identity match. As new identity fragments are obtained, they will be classified and a weight will be assigned to it.

Table 1 shows some examples of identity fragments, the likeliness that this fragment will match multiple natural persons and then the assigned weight.

Table 1 is by no means an exhaustive list of all possible identity fragments and their weights. It should be noted that the likeliness (and thus weight) might and will differ depending on country and culture.

Once PICoCo is satisfied with the classification and confidence of an identity fragment, it will be added to existing identity models.

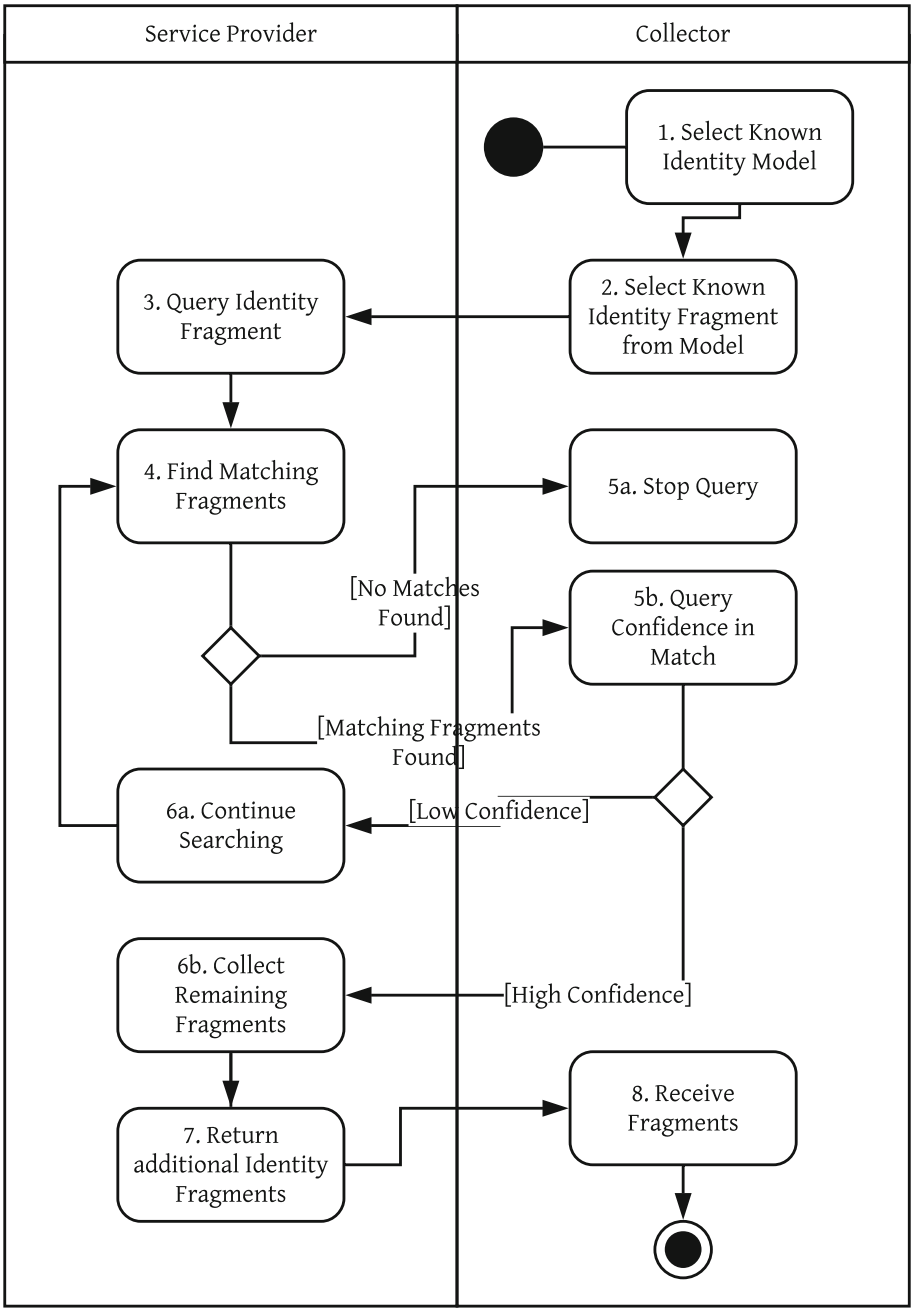

Fig. 3. Process of collecting identity fragments from service provider

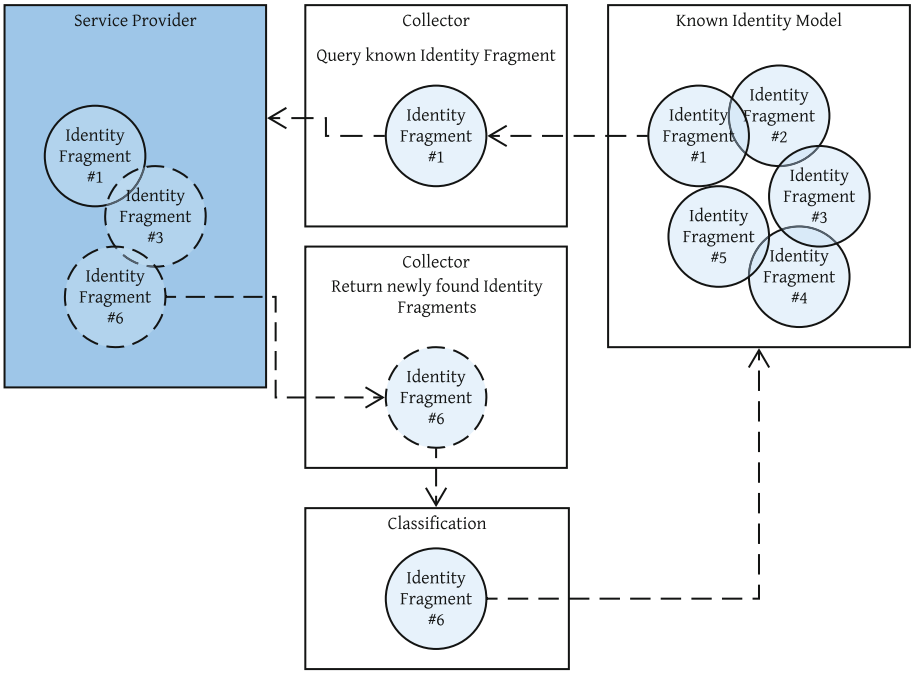

Fig. 4. Collector using identity fragment to identify new fragments

Table 1. Examples of identity fragment weights

Identity fragment	Likeliness of multiple matches	Weight
Nationality	Extremely high	1
Full name	Very high	3
Address	Likely	10
Mobile phone number	Highly unlikely	20
Email address	Highly unlikely	20
USA SSN	Highly unlikely	20
Passport number	Extremely unlikely	30

3.3 Component 3: Collation

Once classified, newly acquired identity fragments will be added to the existing identity model of the queried natural person.

As seen in Sect. 3.1, PICoCo will collect unknown identity fragments from service providers using existing known identity fragments as a means of comparison. PICoCo will use the process outlined in Sect. 3.2 to classify the new incoming identity fragments.

Once collected and classified, the new identity fragment collected will be added to the existing, known, identity model in our data store. All fragments are stored with its weight as well as the service provider that provided us with this new identity fragment.

Finally, all fragments are stored with the date and time it was collected. This will allow PICoCo, in the future, to refer to this when attempting to classify new incoming identity fragments as well as to classify identity fragments are possibly stale or outdated.

3.4 Component 4: Validation

Identity fragments can become stale as they are not updated or revised. We discuss this in Sect. 2.3. Because of this, PICoCo needs to ensure that the identity models that we have created contain the newest and most recent data.

Similar to the collection process discussed in Sect. 3.1, PICoCo will query the known identity providers of all identity fragments in question whether the user has updated or changed this particular fragment. If so, PICoCo will follow the Classfication and Collation process to add this new identity fragment into the identity model.

There might be cases, however, where PICoCo will need to consult with the user on the accuracy of their data. If the confidence of an identity fragment being updated does not meet the required threshold, PICoCo can directly query known users about the accuracy of the data we believe to be correct. Data provided by the user themselves, will carry a very high weight toward the confidence of the accuracy of the identity fragment.

However, this will only be possible if the user has subscribed to the service provided by PICoCo. Not only will PICoCo keep user data up to date, but also provide the user with a list of service providers that has a user's personal information. We discuss this more in the following section.

3.5 Component 5: Discovery

Users should have the right to know which service provider has what information about them. To this end, PICoCo will allow a user to query our identity model of them to discover what is known about them.

As PICoCo is already collecting and collating information about the users, it can easily provide this information to the user. One caveat, however, would be to confidently identify and authenticate the user wishing to discover their identity.

For this process, PICoCo will need to ask a series of questions based on the identity we already know. This is not too unsimilar to certain credit beauraux asking users to identity themselves.

Once the user is confidently identified and authenticated, we can provide the personal information of them we have connected to each service provider that also has this particular identity fragment.

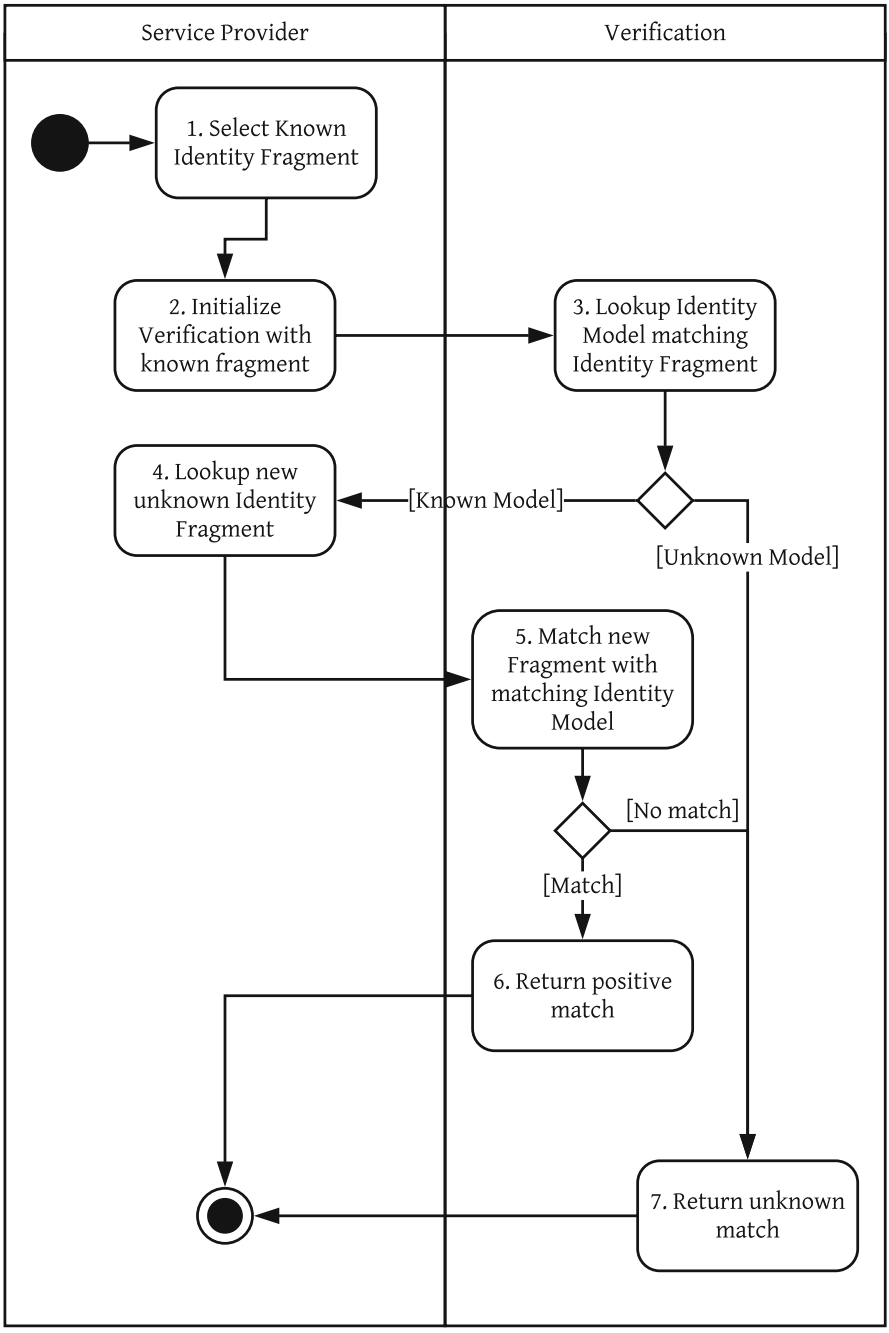

Fig. 5. Verification process from service providers

3.6 Component 6: Verification

Finally, PICoCo can also be used a means to verify new personal information. Service providers should still actively ask their users to maintain up-to-date personal information. However, they can use this model to verify the potential accuracy of the provided information provided. This process is shown in Figs. 5 and 6.

As a user updates their personal information or provides new personal information, the service provider can log a query. They select and provide known personal information (steps 1 & 2) and provide this to the Verifier.

The verifier will look for a matching identity model (step 3). If not none is found, an unknown match is returned (step 7). Otherwise, the service provider will provide the newly acquired personal information to be tested against this known identity model (step 4). PICoCo can then verify whether it believes the queried personal information matches identity fragments that PICoCo has against a known identity model (step 5). If it does, a positive match is returned (step 6), otherwise, PICoCo will return that it is not confident whether data can be linked and verified to an agreed confidence (step 7).

PICoCo's confidence of linking and verification will depend on the type of personal information being verified. PICoCo uses a similar approach as discussed in Sect. 3.2.

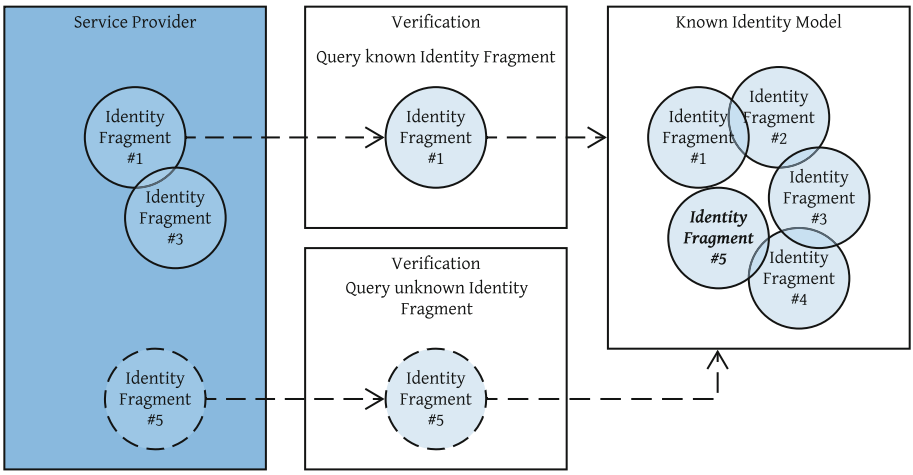

Fig. 6. Verification of identity fragments from service providers

No Information Provided Actively Provided to Service Providers. At this stage, we feel it should be explicitly noted that PICoCo will *never* actively provide or give personal information to service providers! The verification can only be used for actively confident identity models and the verification of additional personal information against known identity fragments.

4 Analysis of PICoCo

In this section we will briefly analyse PICoCo described in previously in Sect. 3. We will analyse it in terms of privacy concerns, accuracy, feasibility, and future work.

4.1 Privacy Concerns

Probably the biggest concern of PICoCo is that of privacy. Essentially the question that can be asked is "How does collating all personal information help privacy?" The simple answer is that the collection and collation of a natural person's personal information, spread across the Internet, will allow better oversight over their current state of affairs.

By allowing users to keep track of where their personal information is being stored and how it is being used, will allow such a person to make informed decisions going forward. Decisions with regard to how they provide and supply their own information to service providers will now be made with a better understanding of such a service provider.

Users who are unfamiliar with the idea of security online and how data online can affect their privacy, can make use of other services that can be provided by PICoCo. One such an approach is described in [1] where a online privacy metric and score is calculated to give the user an idea of their current privacy state.

It should also be noted, that we will under no circumstances force a user or a service provider to make use of this service. However, the advantages of applying such a service to their personal information will be made clear.

The issue of privacy raised by this model is closely linked to that of feasibility, discussed later in Sect. 4.3.

4.2 Accuracy of Personally Identifiable Information

The simple fact that PICoCo relies on other service providers for updated PII should bring the accuracy of said PII into question.

While it is by no means perfectly assured, there can be high level of confidence that personal information on said service providers – especially when compared and collated by other service providers – will be accurate.

Service providers still have to go through their verification process to ensure the accuracy of personal information provided to them. The use of this model is merely there to collate said personal information, as well as provide an overview of personal information to the user. Using this model for verification purposes is merely an advantageous side-effect.

4.3 Feasibility

Questions to be asked:

"Is this model feasible?"
"Will service providers agree to this?"
"Will users agree to this?"

At this point in design, and in our society, this type of model is purely fantastical and slightly offbeat. However, we feel that the investigation of such a model is very necessary. There are multiple benefits that can be gained from the implementation of such a model, but equally as many disadvantages.

However, considering the PKI Model, we use and actively trust central trust authorities – Certification Authorities (CAs). In the PKI model, the entire process relies on the CA. In the same light, we claim that the central agency/agencies for a PICoCo implementation could be seen as a modernised CA.

The answers to the questions at the start of this subsection is probably: "No". But, looking for the future, there might be some use for it. We discuss some possible future work in the following subsection.

4.4 Future Work

The model described in this paper is very abstract at its very least. For the future, we would like to implement an initial working prototype of this model.

The first perceived step would be to look at the service provider side and design a practical integration with existing service providers. Taking into account that different service providers will store personal information differently, we will need to investigate the least-obtrusive means of integrating the collection part of our model into existing systems.

The second would obviously be to then look at implementing the collector itself along with its collator and classifier. This will involve basic principles of data identification and data sorting. Storing the data in the most efficient way will also be important – from a retrieval efficiency perspective as well as storage size perspective.

Finally, we would like to look at how we will allow users to query their personal information. Most importantly would be investigate means of identifying and authenticating the user.

5 Conclusion

In this paper, we investigated identity models and how they are composed of identity fragments consisting of Personally Identifiable Information (PII). Ultimately we introduced and discussed our PICoCo (Personal Information Collection and Collation) Model that collects personal information in such a way to allow it build a digital identity model for a natural person. PICoCo's components were discussed and how they operate together.

Finally, we performed a brief analysis on the model and looked at future work possible from this model.

References

1. Blauw, F.F., von Solms, S.H.: Towards quantifying and defining privacy metrics for online users. In: 2017 IST-Africa Week Conference (IST-Africa), pp. 1–9, May 2017
2. Boardman, R., Sasse, M.A.: Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 583–590. ACM (2004)
3. Council of European Union: What is personal data? https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en
4. Council of European Union: Council Directive 95/46/EC (EU) (1995). http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:31995L0046
5. Council of European Union: Council Regulation (EU) 2016/679 (2016). http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32016R0679
6. Krishnamurthy, B., Wills, C.E.: On the leakage of personally identifiable information via online social networks. In: Proceedings of the 2nd ACM Workshop on Online Social Networks, WOSN 2009, pp. 7–12. ACM, New York (2009). http://doi.acm.org/10.1145/1592665.1592668
7. McCallister, E., Grance, T., Scarfone, K.A.: Guide to protecting the confidentiality of personally identifiable information (PII). Technical report (2010). http://ws680.nist.gov/publication/get_pdf.cfm?pub_id=904990
8. Milberg, S.J., Burke, S.J., Smith, H.J., Kallman, E.A.: Values, personal information privacy, and regulatory approaches. Commun. ACM **38**(12), 65–74 (1995)
9. Norberg, P.A., Horne, D.R., Horne, D.A.: The privacy paradox: personal information disclosure intentions versus behaviors. J. Consum. Aff. **41**(1), 100–126 (2007)
10. Oxford University Press: Definition of natural person. https://en.oxforddictionaries.com/definition/natural_person
11. Phelps, J., Nowak, G., Ferrell, E.: Privacy concerns and consumer willingness to provide personal information. J. Public Policy Mark. **19**(1), 27–41 (2000)