Master Thesis: Open Specification of a user-controlled Web Service for Personal Data

G. Jahn

November 21, 2016

Abstract

Data is the currency of tomorrow. Organizations, whether in the private or public sector, gathering enormous amounts of personal (big) data. This data is harvested and incorporated by these third parties, but were created by individuals and therefore should belong to them. People depending on all their data. Their identity as well as their personality are defined by their personal data. Meanwhile data silo operators are hammering onto these haystacks eager trying to find any kind of correlations worth interpreting, thereby almost inevitably discriminating the rightful owners. To reduce the possibility of discrimination only the least amount of data that is required should be handed over to the third party. Thus the individual has to be in charge of the whole process. A personal data service will empower its user to regain full control over her data and facilitates detailed information on every data flow. To be able to trust such a tool, the user should be able to look inside. Therefore a personal data service has to be open source and developed transparently, which then also allows to self-host it.

Contents

1

Introduction

1.1 Motivation

Nowadays it is rare to find someone who doesn't collect data about some kind of things, particularly humans are the targets of choice for the *Big Data Movement* [1]. But since humans are all individuals, which means we are all distinct from each other - more or less. Some of us are sharing a bit more similarities, but most are much less similar to each other. Since these few similarities are widely shared, they should be less important, because it seems to be intuitively the nature of inflationary occurrence, but instead

the opposite happens to be the case. It allows to determine who is part of a subset with a specific, and therefore shared, attribute and who isn't, in order to relate apparently common known stereo typical patters onto the individuals in that subset just to predict outcomes of the corresponding problem or question. In other words, searching for causation where in best case one might find correlations - or so called *discrimination*, which

> [...] refers to unfair or unequal treatment of people based on membership to a category or a minority, without regard to individual merit. *[2]*

When interacting directly with each other, discrimination of human beings is still a serious issue in our society, but also when humans leveraging computers and algorithms to uncover former unnoticed information in order to include them in their decision making. For example when qualifying for a loan, hiring employees, investigating crimes or renting flats. Approving or denying is all happening based on data about the affected individuals [3], which is nothing but discrimination, just on a much larger scale and with less effort - almost parenthetically. The described phenomenon is originally referred as *Bias in computer systems* [4]. What at first seems to look like machines going rouge on humans, is in fact the *cognitive bias* [5] of the human nature, modeled in machine executable language and made to reveal the patterns their creators were looking for - the *"Inheritance of humanness"* [6] so to say.

In addition to the identity-defining data mentioned above humans have the habit to create more and more data on a daily basis - pro-actively e.g by writing a tweet and passively e.g by allowing the twitter app accessing their current location while submitting the tweet. As a result already tremendous

amounts of data keeps growing bigger and bigger waiting to get harvested, collected, aggregated, analyzed and finally interpreted. The crux here is, the more data being made available [7] to *mine* on, the chances are much higher to isolate data sets, that differ from each other but are coherent in theirselves. Then it is just a matter of the preceded questioning how to distinguish the data set and thereby the related individuals from each other.

So to lower potential discrimination we either need to erase responsible parts from the machines, therefore raising awareness and teaching people about the issue of discrimination are crucial, or we're trying to prevent our data from falling into these data silos. The later will be addressed in this work.

1.2 Purpose & Outcome

From an individual's perspective providing data to third parties might not seem harmful at all. Instead eventually one get improved services in return, e.g. more adequate recommendations and fitting advertisement, or more helpful therapies and more secure environments. That said, though it is a matter of perception what's good and bad, what's harmful and what's an advantage. Computing data to leverage decision making is essentially just science and technology and it's up to the humans how such tools are getting utilized and what purposes they are serving. Hence it should be decided by the data creators, how their data get processed and what parts of them are used.

To tackle the described issue the initial idea here is (1) to equip individuals with the ability to control and maintain their entire data distribution and (2) thus reducing the amount of *potentially discriminatory* [2] attributes leaking

into arbitrary calculations. To do so people need a reliable and trustworthy tool, which assists them in managing all their *personal data* and making them accessible for 3rd parties but under their own conditions. After getting permission granted these data consumers might have the most accurate and reliable one-stop resource to an individuals's data at hand, while urged to respect their privacy at the same time. However this also comes with downsides in terms of security and potential data loss. Elaborating on that and discussing different solutions will be part of the [design process][Design].

The way how to solve the described dilemma is not new. Early days of work done in this field can be dated back to the Mid-2000s where studies were made e.g. about recent developments in the industry or user's concerning about privacy, and the term *Vendor Relationship Management (VRM)* were used initially within the context of user-centric personal data management, which also led into the *ProjectVRM* [8] started by the *Berkman Klein Center for Internet & Society at Hardvard University.* Since then a great amount of effort went into this research area until today, while also commercial products and business models trying to solve certain problems. For instance concepts such as the *Personal Data Store (PDS)* [9] or a *MyData* [10] implementation called *Meeco* [11], which will all be covered in a more detailed way within the following chapter.

The work and research done for this thesis will be the foundation for an *Open Specification*, which by itself is a manual to implement a concept called *Personal Data as a Service.* Important topics like how the architecture will look like, where the actual data can be stored, how to obtain data from the external API or what requirements a user interface for data management need

to satisfy, will be examined. After the thesis will be finished, the majority of core issues should already be addressed and can then get outlined in the specification document. Only then the task to actual implement certain components can begin. The reason for that is, when sensitive subjects especially like people's privacy is at risk, all aspects in question deserve a careful considerations and then get addressed properly. Thus it is indispensable to put adequate effort primarily into the theoretical work. To be clear though, that doesn't mean writing code to test out theories and ideas can't be done during research and specification development. It might even help to spot some flaws and eventually trigger evolvement.

To ensure a great level of trust to this project and the resulting software, it is vital to make the development process fully transparent and encourage people to get involved. Therefore it is required to open source all related software and documents [12] from day one on.

In summary, this document is meant to be the initial step in a development process fabricating a tool to manage all data that defines her identity controlled and administrated by it's owner, and maybe give her a more precise understanding about where her personal information flow and how this might effect her privacy.

1.3 Terminologies

Web Service: TODO

Open Specification: TODO

Big Data: deep learning, neural networks

Personal Data: TODO

Personal Data as a Service (PDaaS): a web service controlled, owned and maybe even hosted by an individual, which provides access to the owner's personal data and offers maintainability as well as permission management.

Personal Data Store: TODO

Vendor Relationship Manager: [13]

Personal Information Management Systems (PIMS): [14]

serverless: TODO https://auth0.com/blog/2016/06/09/what-is-serverless/

Digital Footprints: TODO

Owner: person who controls (and probably hosts) the data service containing her personal data

(Data) Consumer: Third party, external entity requesting data, authorized by the owner to do so

Data Broker(s): entities with commercial interests, that collect, aggregate and analyze information/data of any kind - in this case about human beings - from different sources in order to enrich the data sets, to finally license the resulting corpora to other organisations. [15]

2

Fundamentals

2.1 Personal Data, Ownership and Digital Identity

- *Personal Data* definition
  - general - freely spoken
  - as of EU law (incl citation)
  - as of US law (incl citation)
  - is it just policy/guideline or enforceable too (law/rule)? what relevance/impact have companies *terms and conditions*?
  - EU and USA (since server might be located outside the state or

effective range)

- *Ownership* of personal data
  - who is the owner in what situation or under what circumstances?
  - am I the owner when I was the one who was collecting them? Does it depend whether the resource was public or somewhat private?
  - what will happen with her data service after a person died?
- *Digital Identity*
  - what is a *DI*? and in comparison to *Personal Data*?
  - what is required to make the PDaaS used or seen as a *DI*?

- In the context of this document and all related work, *Personal Data* (or *Personal Information*) is specified as a set of data points (key-value-pairs), that are related to an individual or defining her as such. Combining a subset of these attributes can result in a unique fingerprint as well as values of single attribute can be unique, depending on the context. Not only external imposed attributes, such as social security number, birth or customer ID, are part of an individual's personal data. Also the data created by individual, no matter if pro-active or passively - belongs to her; for instance time series data such as geo-location or blog posts. All personal data solely associating with an individual, holds as well the ownership of the same.

- The european *Data Protection Regulations* defining *Personal Data* as follows: > 'personal data' means any 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; > *[16]*

- The U.S.A. has little legislation on defining and protecting consumer's privacy. At least they have no explicit bills addressing such area [17]. Though some of the existing sectoral laws consist of partially applicable policies and guidelines [18]; most of them addressing specific types of data. In 2015 the White House made an attempt to fill the gap with the *Consumer Privacy Bill of Rights Act*, but to this date it didn't passes the draft state. According to the critics, it lags of concrete enforceable rules consumers can rely on [19]. The draft contains a general definition of *Personal Data*: > "Personal data" means any data that are under the control of a covered entity, not otherwise > generally available to the public through lawful means, and are linked, or as a practical matter > linkable by the covered entity, to a specific individual, or linked to a device that is > associated with or routinely used by an individual, including but not limited to [...] > *[20]*

- followed by a list of concrete data points, e.g. email or postal address, name, social security number and alike. Aside from the legislation with bills, a few third-party organisation can also participate by and add new or overwriting existing rules and policies. Namely for example the *Federal Communications Commission* (FCC), recently releasing *Rules to Protect Broadband Consumer Privacy* including a list of categories of sensitive information [21], which wants *Personally Identifiable Information* (alias Personal Data) to be understood as: > [...] any information

that is linked or linkable to an individual. […] information is > "linked" or "linkable" to an individual if it can be used on its own, in context, or in > combination to identify an individual or to logically associate with other information about a > specific individual. > *[22]*

- Despite minor difference in detail, they all have similar ideas of personal data and their belonging. Even though, the version proposed by EU is almost identical with the definition introduced for the context of this work. Although the FCC's statutory authorities might be somewhat debatable regarding certain topics, the *Communications Act* as a U.S. federal law equips the FCC with power to regulate and legislate.

- Having a common opinion on what data points are belonging to person is the foundation to define a set of rules on how deal with *Personal Data* accordingly. Every business, operating within the EU, is required[1] to provide it's users with a *Privacy Policy*, while e.g. in the U.S. - as mentioned above - only partially and depending on context and data type users must be informed about which and how their data get processed [23].

- A user commonly agrees on the privacy policy, by starting to interact with the author's business, thus every *Privacy Policy* is required to be publicly accessible; e.g. before creating an account. > By clicking Create an account, you agree to our Terms[2] > and that you have read our Data Policy[3], including > our Cookie Use[4]. > *[web_2016_facebooks-landing-page_policy-acknowledgement]*

[1] according to article 12-14 of the *EU General Data Protection Regulation 2016/679*
[2] https://www.facebook.com/legal/terms
[3] https://www.facebook.com/about/privacy
[4] https://www.facebook.com/policies/cookies/

- It can be seen more likely an information notice, that translates and specifies general given law, rather then a contract.
- With such knowledge at hand, it is up to each individual, if the service's benefits are worth sharing some personal data, while simultaneously acquiescing potential downsides concerning the privacy of such data.
- Every entity who is doing so, muss process Personal data according to the law and their *Privacy Policy*. If they policies are violating existing law or the entity effectively goes against the law with their actual doing, penalties might follow. Depending on the level and impact of their infringement in addition the law itself, aside from revising their wrongdoings the entity might have to compensate the affected individuals, pay a fine or get revoked their license.
- Not only privacy laws, but every legal jurisdiction has it's limitations - concerning their territorial nature - which makes legislation not exactly an appropriate tool when it comes to fixing existing issues and strengthen the individual's privacy and rights in a global context like the *world wide web*. If no international agreement is in place [24], only those laws are considered valid and enforcible where the organisation is registered, and maybe the fact where (meaning in which area of jurisdiction) the their servers are located or the data is processed and stored.

Whereas **Ownership** of *Personal Data* is generally addressed in the organisation's *Terms of Service*, which an individual might need to accept in order to establish a (legal) relationship with it's author

the contents (of ToS) is not against any appliable law; or the regarding issues are lacking of any legislative addressing.

- Since the U.S. law barely handles consumer privacy, it also touches just briefly on ownership of data and in a rather generic manner.
- A **Digital Identity** goes a step further by not only representing and associating a living human being, but also providing an additional level of authenticity verification/insurance.
- NOTE: maybe the dif goes as follows - personal data becomes a *DI* only when the data set holds enough attributes or specific attributes to be unique in the given context, and therefore allowing to identify associated individual (??); not really, same goes for personal data
- all digital data about, related to and created by an individual, that would also identify this person as the rightful owner and physical counterpart. It can also be seen as an avatar in the digital world or as the digital part of a human's identity. [25]
- identity defining data (e.g. history of personal ID card)
- with such a system a human being is represented by a non-physical abstraction of herself. Which essentially is a list of attributes, that are at least for legal and civil administration purposes important. Their values in total are unique and representing the corresponding human. Certain attributes hold unique values within it's own context, for example the *social security number*.
- Thus it's not necessary to know the values of all attributes in order to identify it's owner

- therefore its imported to not see it as a reduction of a living individual to some bits and bytes

[26]

- ownership? (https://united-kingdom.taylorwessing.com/globaldatahub/article_big_data_
  https://lifehacker.com/you-dont-own-your-data-1556088120
- https://www.eff.org/deeplinks/2015/07/big-tech-does-not-speak-internet

2.2 Personal Data in the context of the Big Data Movement

- big data itself initially can be seen as a *huge blob of data* containing more or less structured data sets [27], whose size might have exceeded the capabilities of retrieving certain information almost only by hand. Such high data haystacks usually come along with new challenges in logistic and resource management, when information retrieval needs to get automated on a large scale [28]. Theses practices are commonly referred to *Big Data (Analysis)* including distributed computing and machine learning.

- Big Data, or to be more precise, collecting and analyzing big data, serves the prior purpose to extract useful information, which on the other hand depends on what was the opening question about, but also what data sets the corpus is containing.

- At first, (A) formalizing question(s) that the results have to answer.

Secondly, (B) deciding what data is needed and appropriate and then start collecting. Third, (C) designing data models accordingly and correlate with the data (D) next, analyse and interpret the results. (E) last but not least, make business decisions based und the analyses ([29] Fig. 3).

- since quite a few businesses (in terms of purpose or intention) are based around the concept of customers, which are generally somewhat entities consisting of at least one human being, personal data takes a major part in what *Big Data* can be about. In the context of this thesis, these entities are individuals with a unique identity. And to understand the behaviour, decision making and needs of her customers a vendor, who owns the business, needs to know as much as possible about them, when she wants to know what changes she needs to address in order to move towards the most lucrative business.

- personal data and information are reflecting all this knowledge. It starts with profile data, such as gender, age, residency or income, goes on with time series events like geo-location changes, or web search history and goes all the way up to health data and self-created content like *Tweets*[5] or videos.

- all these classes of personal data hold a major share[6] in the field of data analytics (TODO: find statistics showing shares of data

[5]public massages published by an account on twitter.com, which will be displayed in the timeline of all her subscribers and also might contain additional types of content like images, links or video

[6]it doesn't matter whether an individual or just someone on behalf of an organisation spend money for something. at the end of the day, they are all humans on this planet and in a capitalistic oriented world money needs to flow and profits needs to be maximized. So to know where it will flow or why it will flow in a certain direction it is crucial to know everything about it's decision maker - the humans on this planet.

types/classes/categories, [30] [31])

- but, depending on the specific attributes, they might be not that easy to acquire. in general most businesses obtain data from within their own platforms. some data might be in the user's rang of control (e.g. customer or profile data), but most of the data comes from interacting directly (content creation, inputs) or indirectly (transactions, meta information). the level of sensitivity is mainly based on the purpose of the platform (benefit for the user) and what is the provider's demand from the users commitment (e.g. required inputs or usage requires access to location)

- from a technical perspective collecting passively created data is as simple as integrating logging mechanisms in the program logic. since the industry moved towards the cloud[7] most scenarios utilized server-client architectures. Furthermore the *always-on* philosophy evolved to an imperative state. standalone software is starting to call the author's servers from time to time, just to make sure the user behaves properly. For browsers it was already a common narrative to make here and then requests to the server - still preventable though, but when it comes to native mobile apps it is almost impossible [32] to notice such behaviour and therefore preventing apps from doing so.

- these architectural developments were inducing the gathering of potentially useful information from all over the system on a large scale [33]. Logging events, caused by the user's interactions, on the client, which

[7] side note - one might come to the conclusion, that only the trend towards the *cloud* made it actually possible to collect to such an extent we are all observing these days, because standalone software should not necessarily require internet connection and therefore the vendors had no way to gather information whatsoever

then get forwarded to backend servers. Or keeping track of all kinds of transactions, which is done directly in the backend. Before running together in a designated place, all these collected chucks of data (TODO or "data points") are getting enriched with meta information. Finally get stored and probably never removed again - all for later analyses.

- The mindset in the *Big Data Community* is grounded on the basic assumption of *more data is more helpful*, which already is emphasised by the often-cited concept of the three *Vs* (Volume, Velocity, Variety) [34]. which is not entirely wrong, because it lies in the nature of pattern and correlation discovery, to provide increasing quality results [35], while enriching the overall data with more precise data sets. But when new technologies are emerging, questioning the downsides and possible negative mid- or long-term impacts are typically not very likely to be a high priority. The focus lies on e.g. trying to to reach and eventually breach boundaries while beginning to evolve. So non-technical aspects such as privacy and security awareness doesn't come in naturally, instead a wider range of research needs to be done alongside the evolution process and the increasing adoption rate in order to uncover such issues. Only then they can addressed properly on different levels - technical, political as well as social. So that the *Big Data Community* itself is able to evolve, too. All in all it's a balancing act between respecting the user's privacy and having enough data at hand to satisfy the initial questioning with the computed results. Therefore people working in such contexts need to have advanced domain knowledge, be aware of any downsides or pitfalls and need to be sensible about the ramifications of their approaches and doings. Such improvements are already

happening, not only originating from the field's forward thinkers [36], but also advocated by governments, consumer rights organisations and even leading Tech-Companies start trying to do better [37] [38] [39] - as discussed in the section [TODO see personal data as of the law],

- earlier in the text a difference was made between actively created and passively created data
- based on that one could say *profile/account data* is actively created, because it got into the system by the user's actively made decision to insert these information into a form and submit it - for whatever reason. whereas detecting the user's current location and adding this information to the submitted form is *meta data*
- of cause, it is debatable whether these kind of data belongs, in the sense of being the rightful owner, to the user or to the author or owner of the software containing the code that effectively created the data.
- maybe personal data is every data/information whose creation (or digital existence) is a direct result of user interaction/engagement?
- lets have a look into what the rule book says about that –> next topic (law)

2.3 Personal Data as a Product

- *Big Data Analytics* by itself just comprises a structured and technical-aided procedure, serving the purpose of finding invisible information, that might be helpful to make (right) (business) decisions. Though, if

one would ask data collectors about their motivation, most likely the answer would be something along the lines of PR phrasing like *"We want to have a better understanding of our customers"*. But to do what exactly? To predict what might be the next thing I am supposed to buy Or what things I probably would like to consume but most certainly not yet know of?

- Let's take a look at some examples. An advertising service uses tracking data for targeted advertising. The more information they have about an individual, the more accurate decisions they are able to make about what ads are the ones the individual most likely will click on and disclose with a successful purchase. As a result this makes the placed advertisement more valuable for ad service and therefore more expensive to the advertisers, because of a high precision. Or a streaming provider's content recommendation is also based on heavy user profiling done by looking at her consumption history, tracked platform interactions and probably many more vectors. Another example is *Google Traffic* [40] [41], a service, integrated as a feature in *Google Maps*, which is Google's web mapping service. *Google Traffic* visualises real-time traffic conditions, when using *Maps* as a navigation assistant, to provide the user with a selection of possible paths, but enriched with duration, that takes such conditions into account. The data, required to offer these information, is supplied by mobile devices, constantly sending GPS coordinates with a timestamp into Google's infrastructure. This, however, only is made possible, because Google's services are widely used in addition to the fact that the majority of mobile devices [42] is driven by Android, an mobile operating system

developed by Google, that deeply integrates with it's services. For this case the same assertion can be made - the more constantly streaming geo-location data, the more precise the information are about traffic conditions. Since this information demands the real-time aspect, adding time to the equation, add a other dimension of complexity to problem.

- while the impact on our society of this first example group might be doubtable, a change of perspective opens up a different range of application areas. Such as
  - planing and managing human resources for situations, like e.g. big events or emergency situations where attendees might need some help [43]
  - predicting infrastructure workloads [TODO http://ieeexplore.ieee.org/document/7336
  - making more accurate diagnostics to improve their therapy [44]
  - finding patters in climate changes, which otherwise wouldn't be detected [45].
- Through all these examples, some of them might not necessarily founded on personal data, whereas others primarily depend on them and yet others only implicitly rely on data collected from individuals. As always, it depends on the purpose - also known as *business model* - but it seems to be consensual, that it all comes down to improving and enhancing the collector's product in order to satisfy the customers - and that on the other hand depends on what is meant to be the product and who is seen as customers.
- Putting a top 10 list of industries using utilizing *Big Data* [46] right

next to visualization showing categories of personal data targeted by data collectors

[47], at least 7[8] of these industries can be identified as data collectors, whereas less then a half[9] are taking part of being a *Data Broker*, but almost all of them are using people's personal data, whether collected by themselves or acquired from *Data Broker*.

- At this point it's save to say, that *Personal Data* is either seen directly as a product, especially from a Dater Broker's point of view, or indirectly due to it's essential part in *Big Data* practices. The former generates direct revenue by selling these data and the latter might affect a business's product quality in a positive manner and thereby increasing revenue as well.

- At the end it all comes down to understanding the human being and why she behaves as she does. The challenge is not only to compute certain motives but rather concluding to the right ones. When analyzing computed results with the corresponding data models and trying to conclude, it is important to keep in mind, that correlation is by far no proof of causation.

- individuals then get in role of selling/offering it's own data to those who were previously collecting them

[8]Banking and Securities; Communication, Media & Entertainment; Healthcare Providers; Government; Insurance; Retail & Wholesale Trade; Energy & Utilities

[9]Banking and Securities; Communication, Media & Entertainment; Insurance; Energy & Utilities

2.4 Use Cases (NOTE: maybe move to 100_introduction)

- package shipment after buying sth online
- social network accessing arbitrary profile data
- making an online purchase
- credibility (applying for a loan) validation by a certain financial institution: accessing arbitrary data
- patient/health record
- care (movement) data

2.5 Related Work

- much more research since the data mining, big data, deep learning

Different terms, same meaning:

- Personal Agent
- Personal Data Vault
- Personal Data Store

2.5.1 Research

- openPDS/safeAnswer [http://openpds.media.mit.edu/]
- TAS3 aka ZXID aka Synergetics (lead arch Sampo Kellomäki also Co-Authored openPDS papers)
- Higgins [https://www.eclipse.org/higgins/]

- Hub-of-All-Things [http://hubofallthings.com/what-is-the-hat/]
- ownyourinfo [http://www.ownyourinfo.com]
- PAGORA [http://www.paoga.com]
- PRIME/PrimeLife [https://www.prime-project.eu, http://primelife.ercim.eu/]
- databox.me (reference implementation w/ the "solid" framework)
- Microsoft HealthVault
- Industrial Data Space (german research project mainly driven by Fraunhofer Institute)
- Polis (greek research project from 2008) [http://polis.ee.duth.gr/Polis/index.php]

2.5.2 Organisations

- Kantara Initiative (former "Liberty Alliance") [https://kantarainitiative.org/]
- Open Identity Exchange [http://openidentityexchange.org/resources/white-papers/]
- Qiy Foundation [https://www.qiyfoundation.org/]

2.5.3 Commercial Products

- MyData [https://mydatafi.wordpress.com/]
- Meeco (killing the ad provider middle man) [https://meeco.me/how-it-works.html]
- RESPECT network [https://www.respectnetwork.com/]
- aWise AEGIS [http://www.ewise.com/aegis]

2.6 Standards and Specifications

- http(s)
- all the *Semantic Web* stuff
- Container/App spec
- JWT
- oAuth (?)
- JSON
- REST
- GraphQL

3

Core Principles

NOTE: here we discuss a variety of possibilities –> conceptual work

3.1 Data Ownership

- user-centric, full control

3.2 Identity Verification

- maybe go with a Signing/verifying Authority (aka CA)
  - do I trust the gov or certain companies more? Which interests do these Role/Stakeholder have?
  - revoking the cert which provides the authenticity of the individual's digital identity should only be possible with a two-factor secret. One part of this secret is owned by the CA and the other half has the individual behind the personal API
- TODO: look into
  - PKI
  - ePerso
  - E-Post/de-mail
- Authentication

3.3 Authorisation (remove in favour of data access?)

- NOTE: does not mean this tool authenticates it's owner against third party platforms like OpenID does. but it could play the role of the 2n factor in a multi-factor authentication process (if the mobile-device-architecture was chosen)
- refers primarily to the process of a data consumer (third party, which needs the data for whatever reason) verifies her admission to request

3.4 Authentic Data

- is this data (in this case identity) certified or not (results in higher value)

3.5 Supervised Data Access

- pure/plain data request/resonse
- remote computation/execution (assuming there is no client for the consumer) like https://webtask.io/

3.6 Encapsulation

- containerization (coreos, rkt, mirageos aka unikernal)

3.7 Open Development

- which and why open standards
- why open source
- collaborative transparent development
- Hosting & Administration
  - DYI
  - Usability

4

Requirements

4.1 User Interaction

- as effortless as possible

4.2 Management & Organisation

4.3 Administration

5

Design

5.1 Architecture

- showing possible directions, e.g.:
  - cloud or local storage
  - which components can go where
  - remote execution, to prevent data from leaving the system

5.1.1 Overview

- distributed architecture (e.g. notification/queue server + mobile device for persistence and administration)

5.1.2 Components

5.1.3 Plugins

- but for what? and not harm security at the same time? maybe just for data types and admin UI to display analytical (time based) data in other ways
- what about extensions (see iOS 10) to let other apps consume data; only on a mobile device and only if the data is stored there

5.2 Data

- keep in mind to make it all somehow extendible, e.g. by using and storing corresponding schemas
- NOTE: step numbers marked with a * are somehow tasks which are happening in the background and don't require any user interaction

5.2.1 Modelling

5.2.2 Categories (or Classes)

5.2.3 Types

5.2.4 Persistence

- database requirements

5.2.5 Access & Permission

- data needs to have an expiration date

IF01 - Authorizing a consumer to request certain data

1) owner creates a new endpoint URI (like *pdaas.ownersdomain.tld/e/consumer-name*) within the *management user interface*
2) owner passes this URI on to the *consumer*, e.g. through submitting a form or using any arbitrary, eventually insecure channel 3*) consumer need to call this URI for the fist time to verify it's authenticity
3) owner then get's a notification which asks her for permissions to access certain data under the listed conditions 5*) consumer will be informed about the outcome of the owner's decision (NOTE: alongside with some details? how do they look like? XXX need to be in the spec)

5.2.6 Consumption (data inflow)

- how data will get into the system
- hwo is the user able to do that, and how does it works

5.2.6.1 Manually

5.2.6.2 Automatically

5.2.7 Emission (data outflow)

- depending on what category of data, they might need to get anonymized somehow before they leave the system
- oAuth (1.0a and 2) requires consumers to register upfront. Since the current flow indicates that the initial step is done by the owner, that would cause an overhead in user interactions. Although the owner already *authorized* the consumer simply by submitting a unique URI (`pdaas-server.tld/register?crt=CONSUMER_REGISTER_TOKEN`), of which the `crt` is considered private. Even though the registration provides the consumer with mandatory information such as a consumer identifier (v1: `oauth_consumer_key`, v2: `client_id`) and, depending on the client type, a secret (see https://tools.ietf.org/html/rfc6749#section-2), this process it is not part the specification (https://oauth.net/core/1.0a/#rfc.section.4.2, https://tools.ietf.org/html/rfc6749#section-2). This enables the possibility of integrating oAuth into the consumer registration flow

by using the `CONSUMER_REGISTRATION_TOKEN` as oAuth's *client identifier*. The lack of credentials (v1: `auth_consumer_secret`, v2: `client_secret`) would require transferring the consumer identifier done over a secure channel (e.g. TLS). That would leave *oAuth2* as the version of choice, since it relies on *HTTPS* adn therefore makes the *secret* optional. Where on the other side oAuth 1.0a requires a *secret* to create a signature in order to support insecure connections..

- A general and URI for 3rd parties to register (aka requesting authentication) would raise the issue of dealing with spam request and how to distinct these from the actual ones.

5.2.8 History

- data versioning
- access logs

5.3 Interfaces

5.3.1 Internal

- UI for Management & Administration

5.3.2 External

- should there be a way to somehow request information about what data is available/queryable, or would this be result in spam/crawler and security issues (also a question for the topic of permissions/sensibility level of certain data)
- certain types of requests, depending on expire date:
  - "ask me any time"
  - "allowed until further notice"
  - one-time permission (but respecting certain http error codes and possible timeout - that might not count)

6

Specification

- what does *open* in Open Specification even mean?

6.1 Processes (TODO: find another word; “Protocol flows”?)

6.2 Application Programming Interfaces

6.3 Graphical User Interfaces

6.4 Security

- the downside of having not just parts of the personal data in different places (which is currently the common way to store), is in case of security breach, it would increase the possible damage by an exponential rate Thereby all data is exposed at once, instead of not just the parts which a single service has stored
- does it matter from what origin the data request was made? how to check that? is the requester's server domain in the http header? eventually there is no way to check that, so me might need to go with request logging and trying to detect abnormal behaviour
- is the consumer able to call the access request URI repeatedly and any time? (meaning will this be stateless or stateful?)
- initial consumer registration would be done on a common and valid https:443 CA-certified connection. after transferring their cert to them as a response, all subsequent calls

need to go to their own endpoint, defined as subdomains like `consumer-name.owners-notification-server.tld`

6.4.1 Environment

6.4.2 Transport

- communication between internal components *must* be done in https only, but which ciphers? eventually even http/2?

6.4.3 Storage

- documents based DB instead of Relational DBS, because of structure/model flexibility

6.4.4 Authentication

- how should consumer authenticate?

6.5 Recommendations

6.5.1 Software Dependencies

6.5.2 Host Environment

7

Conclusion

7.1 Ethical & Social Impact (TODO: or "Relevance")

7.2 Business Models & Monetisation

- possible resulting direct or indirect business models
- owner might want to sell her data, only under her conditions. therefore some kind of infrastructure and process is required (such as payment transfer, data anonymization, market place to offer data)

7.3 Challenges

- adoption rate of such technology

7.4 Solutions

7.5 Attack Scenarios

7.6 Future Work

- maybe enable the tool to play the role of an own OpenID provider?
- going one step further and train machine (predictor) by our self with our own data (https://www.technologyreview.com/s/514356/stephen-wolfram-on-personal-analytics/)

7.7 Summary

- main focus
- unique features
- technology stack & standards
- resources
- the tool might be not a bulletproof vest, but

[1] "Big data privacy international." [Online]. Available: https://www.

privacyinternational.org/node/8. [Accessed: 15-Nov-2016]

[2] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 560–568 [Online]. Available: http://dl.acm.org/citation.cfm?id=1401959. [Accessed: 03-Nov-2016]

[3] S. Spiekermann, *Ethical IT Innovation: A Value-Based System Design Approach*. CRC Press; Taylor & Francis Group, LLC, 2015, pp. 66–72 [Online]. Available: https://www.crcpress.com/Ethical-IT-Innovation-A-Value-Based-System-Design-Approach/Spiekermann/p/book/9781482226355

[4] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems (TOIS)*, vol. 14, no. 3, pp. 330–347, 1996 [Online]. Available: http://dl.acm.org/citation.cfm?id=230561. [Accessed: 07-Nov-2016]

[5] "Cognitive bias," *Wikipedia*, Oct-2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Cognitive_bias&oldid=742803386. [Accessed: 08-Nov-2016]

[6] R. Lemov, "Why big data is actually small, personal and very human. Aeon essays," 16-Jun-2016. [Online]. Available: https://aeon.co/essays/why-big-data-is-actually-small-personal-and-very-human. [Accessed: 17-Nov-2016]

[7] A. Dewes, "C3TV - Say hi to your new boss: How algorithms might soon control our lives." 29-Dec-2015. [Online]. Available: https://media.ccc.de/v/32c3-7482-say_hi_to_your_new_boss_how_

algorithms_might_soon_control_our_lives#video&t=1538. [Accessed: 03-Nov-2016]

[8] "ProjectVRM - about. ProjectVRM," 25-Feb-2010. [Online]. Available: https://blogs.harvard.edu/vrm/about/. [Accessed: 09-Nov-2016]

[9] Tom Kirkham, Sandra Winfield, Serge Ravet, and S. Kellomaki, "The personal data store approach to personal data security," *IEEE Security & Privacy*, vol. 11, no. 5, pp. 12–19, 2013.

[10] A. Poikola, K. Kuikkaniemi, and H. Honko, "MyData – a nordic model for human-centered personal data management and processing," pp. 1–12, 2014 [Online]. Available: https://www.lvm.fi/documents/20181/859937/MyData-nordic-model/2e9b4eb0-68d7-463b-9460-821493449a63. [Accessed: 10-Nov-2016]

[11] "Meeco how it works." [Online]. Available: https://meeco.me/how-it-works.html. [Accessed: 09-Nov-2016]

[12] "Open specification of the concept called personal data as a service (pdaas). GitHub." [Online]. Available: https://github.com/lucendio/pdaas_spec. [Accessed: 11-Nov-2016]

[13] "ProjectVRM wiki - about VRM." [Online]. Available: https://cyber.harvard.edu/projectvrm/Main_Page#About_VRM. [Accessed: 11-Nov-2016]

[14] "ProjectVRM wiki - list of personal information management systems." [Online]. Available: https://cyber.harvard.edu/projectvrm/VRM_Development_Work#Personal_Information_Management_Systems_

.28PIMS.29. [Accessed: 11-Nov-2016]

[15] F. T. C. USA, "Data brokers," May 2014 [Online]. Available: https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-re 140527databrokerreport.pdf. [Accessed: 17-Nov-2016]

[16] *General data protection regulation.* 2016, p. L 119/33 [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

[17] Wikipedia, "Information privacy law," 13-Nov-2016. [Online]. Available: https://en.wikipedia.org/wiki/Information_privacy_law#United_States. [Accessed: 20-Nov-2016]

[18] I. J. (Loeb & Loeb), "PLC - data protection in the united states: Overview," 01-Jul-2013. [Online]. Available: http://us.practicallaw.com/6-502-0467. [Accessed: 20-Nov-2016]

[19] A. Wilhelm, "White house drops 'consumer privacy bill of rights act' draft. TechCrunch," 27-Feb-2015. [Online]. Available: http://social.techcrunch.com/2015/02/27/white-house-drops-consumer-privacy-bill-of-rights-act-draft/. [Accessed: 20-Nov-2016]

[20] *Administration discussion draft: Consumer privacy bill of rights act of 2015.* 2015 [Online]. Available: https://www.whitehouse.gov/sites/default/files/omb/legislative/letters/cpbr-act-of-2015-discussion-draft.pdf

[21] *Report and order.* 2016 [Online]. Available: https://transition.fcc.gov/Daily_Releases/Daily_Business/2016/db1103/FCC-16-148A1.pdf.

[Accessed: 20-Nov-2016]

[22] *Notice of proposed rulemaking*. 2016 [Online]. Available: https://apps.fcc.gov/edocs_public/attachmatch/FCC-16-39A1.pdf. [Accessed: 20-Nov-2016]

[23] "Privacy policies are mandatory by law," 23-Oct-2016. [Online]. Available: https://termsfeed.com/blog/privacy-policy-mandatory-law/. [Accessed: 20-Nov-2016]

[24] "International privacy standards," 29-Sep-2016. [Online]. Available: https://www.eff.org/issues/international-privacy-standards. [Accessed: 20-Nov-2016]

[25] J. Rose, O. Rehse, and B. Röber, "The value of our digital identity," *Boston Cons. Gr*, 2012 [Online]. Available: https://www.libertyglobal.com/PDF/public-policy/The-Value-of-Our-Digital-Identity.pdf

[26] E. O. of the US President, "Big data: Seizing opportunities, preserving values," The White House, May 2014 [Online]. Available: https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

[27] "Big data n." [Online]. Available: http://www.oed.com/view/Entry/18833#eid301162177. [Accessed: 11-Nov-2016]

[28] Wikipedia, "Big data," 11-Nov-2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Big_data&oldid=748964100. [Accessed: 11-Nov-2016]

[29] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data

analytics: A survey," *Journal of Big Data*, vol. 2, no. 1, p. 21, Oct. 2015 [Online]. Available: http://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0030-3. [Accessed: 13-Nov-2016]

[30] O. R. Zaïane, *Principles of knowledge discovery in databases*. 1999, pp. 1–2 [Online]. Available: https://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/. [Accessed: 13-Nov-2016]

[31] "Big data collection collides with privacy concerns, analysts say. PCWorld," 10-Feb-2013. [Online]. Available: http://www.pcworld.com/article/2027789/big-data-collection-collides-with-privacy-concerns-analysts-say.html. [Accessed: 15-Nov-2016]

[32] "Answers.io. Answers." [Online]. Available: https://answers.io/answers. [Accessed: 14-Nov-2016]

[33] A. L. Burgelman, N. L. Burgelman, and NGDATA, "Attention, big data enthusiasts: Here's what you shouldn't ignore. WIRED." [Online]. Available: https://www.wired.com/insights/2013/02/attention-big-data-enthusiasts-heres-what-you-shou
[Accessed: 15-Nov-2016]

[34] D. Laney, "3D data management: Controlling data volume, velocity, and variety," META Group, February 2001 [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-pdf

[35] M. Hilbert, "Big data for development: A review of promises and challenges," *Development Policy Review*, vol. 34, no. 1, pp. 135–174, December 2015 [Online]. Available: http://dx.doi.org/10.1111/dpr.12142

[36] N. Davis Kho, "The state of big data," 24-Feb-2016. [Online].

Available: http://www.econtentmag.com/Articles/Editorial/Feature/The-State-of-Big-Data-108666.htm. [Accessed: 18-Nov-2016]

[37] T. C. (Apple's CEO), "A message to our customers. Customer letter," 16-Feb-2016. [Online]. Available: http://www.apple.com/customer-letter/. [Accessed: 18-Nov-2016]

[38] M. Green, "What is differential privacy? A few thoughts on cryptographic engineering," 15-Jun-2016. [Online]. Available: https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/. [Accessed: 18-Nov-2016]

[39] B. Budington, "WhatsApp rolls out end-to-end encryption to its over one billion users," 07-Apr-2016. [Online]. Available: https://www.eff.org/deeplinks/2016/04/whatsapp-rolls-out-end-end-encryption-its-1bn-users. [Accessed: 18-Nov-2016]

[40] "Stuck in traffic? Insights from googlers into our products, technology, and the google culture," 28-Feb-2007. [Online]. Available: https://googleblog.blogspot.com/2007/02/stuck-in-traffic.html. [Accessed: 18-Nov-2016]

[41] Wikipedia, "Google traffic," 25-Oct-2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Google_Traffic&oldid=746200591. [Accessed: 18-Nov-2016]

[42] "Global mobile OS market share." [Online]. Available: https://www.statista.com/statistics/266136/global-market-share-held-by-smartphone-operating-systems/. [Accessed: 18-Nov-2016]

[43] J. Ao, P. Zhang, and Y. Cao, "Estimating the Locations of Emergency

Events from Twitter Streams," *Procedia Computer Science*, vol. 31, pp. 731–739, 2014 [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1877050914004980. [Accessed: 05-Nov-2016]

[44] G. Palem, "The Practice of Predictive Analytics in Healthcare," *ResearchGate*, Apr. 2013 [Online]. Available: https://www.researchgate.net/publication/236336250_The_Practice_of_Predictive_Analytics_in_Healthcare. [Accessed: 05-Nov-2016]

[45] N. Burger, B. Ghosh-Dastidar, A. Grant, G. Joseph, T. Ruder, O. Tchakeva, and Q. Wodon, "Data Collection for the Study on Climate Change and Migration in the MENA Region," 2014 [Online]. Available: https://mpra.ub.uni-muenchen.de/56929/. [Accessed: 04-Nov-2016]

[46] M. Gaitho, "Applications of big data in 10 industry verticals," 20-Oct-2015. [Online]. Available: https://www.simplilearn.com/big-data-applications-in-industries-article. [Accessed: 19-Nov-2016]

[47] F. T. C. USA, "Personal data ecosystem," *Protecting Consumer Privacy in an Era of Rapid Change - Recommendations for Business and Policymakers - FTC Report*, March-2012. [Online]. Available: https://www.ftc.gov/sites/default/files/documents/public_events/exploring-privacy-roundtable-series/personaldataecosystem.pdf. [Accessed: 17-Nov-2016]