D4.5

User Categorisation of Digital Footprint – v2

V1.0 / 2015-12-13

Tom Seymoens (iMinds), Rob Heyman (iMinds), Jonas Breuer (iMinds), Laurence Claeys (iMinds), Jo Pierson (iMinds)

This deliverable is an update of D4.2 in which we explore how users evaluate their data in relation to information disclosure. The central question which we want to answer is the following: In what way do people differ about how they deem different types of online information to be private and/or sensitive and does their attitude change when confronted with a multitude of institutional actors reasoning on their data? The results presented below are the product of both quantitative and qualitative analyses of user attitudes and behaviour towards personal information sharing on the internet.

Project acronym	USEMP
Full title	User Empowerment for Enhanced Online Presence Management
Grant agreement number	611596
Funding scheme	Specific Targeted Research Project (STREP)
Work program topic	Objective ICT-2013.1.7 Future Internet Research Experimentation
Project start date	2013-10-01
Project Duration	36 months

Workpackage	WP4
Deliverable lead org.	iMinds
Deliverable type	Report
Authors	Tom Seymoens (iMinds) Rob Heyman (iMinds) Jonas Breuer (iMinds) Laurence Claeys (iMinds) Jo Pierson (iMinds)
Reviewers	Giorgos Petkos (CERTH) Hervé Le Borgne (CEA)
Version	1.4
Status	Draft | PMB Final Draft | **Final**
Dissemination level	**PU: Public**, PP: Restricted Program; RE: Restricted Group; CO: Confidential
Due date	2016-01-01
Delivery date	2016-02-21

Version	**Changes**
V1.0	Table Of Contents + Abstract
V1.2	Introduction update + quantitative research track
V1.3	Qualitative research track
V1.4	Results and finalization

Table of Contents

1. Introduction

The overall goal of WP 4 in the USEMP project is to enhance the understanding of how users apply social platforms in their everyday life and how they evaluate the disclosure of different types of personal information. The work presented in the deliverable at hand specifically revolves around task 4.2. This task aims at gathering greater insights into the on-going data disclosure practices on Online Social Networks (OSNs) and how the users of these platforms evaluate and distinguish between different types of personal information. To this extent it takes into account users' claimed and actual data disclosing behaviour and their attitudes and measured awareness towards how social platforms and third parties reason on their data.

The predecessor of this report is D4.2: User Categorization of Digital Footprint – v1, made public in September 2014. In D4.2 we took a first glance on data sharing practices through the means of an extensive literature study with a focus on which data is generally disclosed, the reasons for information disclosure and retention and how this has evolved over the years. After the publication of this deliverable, the European commission released a highly relevant publication on data protection (European Commission, 2015). We cannot ignore this report in light of the task at hand, so we will start this deliverable in the next chapter with a small overview of its most relevant results. In the second section of the chapter we will present an overview of a categorization of different types of personal information that social platform users disclose and how this data creates value for the social platform and third parties. Finally, we will provide a state of the art overview of data extraction techniques and what is technically possible to infer, based on deliverable 6.1. In chapter 3, the research performed is linked to the work done in other work packages of the USEMP project.

Building further upon the literature study of the second chapter and D4.2, we conducted a quantitative and qualitative research track. The quantitative track enabled us to gather insights on what users claim to find sensitive and private information. The setup permitted us as well to compare actual information disclosure with perceived information disclosure. The study took place in May and June of 2015 with 182 respondents, the majority of which were Swedish and Belgian citizens. The composition, methodology and results of this study are described in detail in chapter 4.

To gather more insights into individual perspectives on the matter of online information disclosure we organized a total of 21 qualitative interviews. Here we wanted to explore how individual users evaluated different types of data that online social platforms endorse to disclose. In 14 of these sessions, we also probed our respondents to get their attitudes towards different types of information collection and five different institutional actors for whom it could be beneficial to reason on user data. To see the full scope of the study, please refer yourself to chapter 5 of the current deliverable.

Figure 1 presents the the different research tracks and how they each add to a greater understanding of online information disclosure.

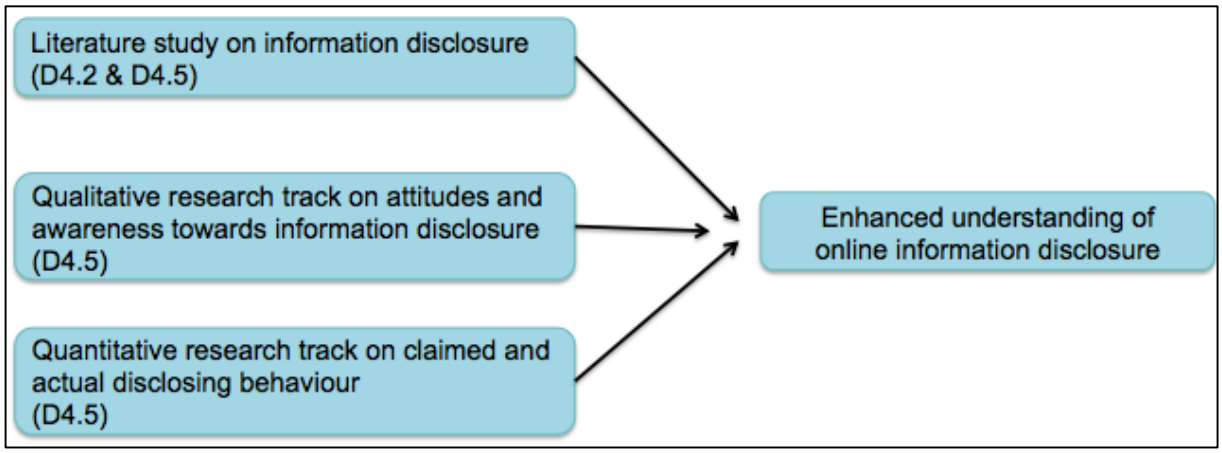


Figure 1: Setup of mixed method research on online information disclosure

2. Literature Study

2.1. Short Summary of the Eurobarometer on Data Protection

A recent version of the Special Eurobarometer, a series of reports issued by the European Commission focusing on specific themes, revolves around data protection (European Commission, 2015). For this edition they conducted a survey with 27,980 respondents stemming from the 28 member states of the European Union in February and March of 2015. The report approaches data protection from a plethora of angles: control, disclosure, rights and protections, management by other parties, etc. Below we highlight some reported results that are relevant to our research.

2.1.1. Concerns about Data Collection

The monitoring of everyday activities online

With respect to monitoring of everyday activities on the Internet (explained in the survey as browsing, downloading files, accessing online content), 45% of the respondents say they are concerned about the recording of their everyday activities, of which 13% say they are very concerned and 32% fairly concerned. 36% are not concerned about the matter, whereas 17% (almost one out of five) say this is not applicable to them. For our research it's interesting to notice that the Swedish population were least concerned about this (only 25% claimed to be concerned), whereas the Belgian population with 55% nested itself in the top three countries that expressed concern (after UK and Ireland).

Government agencies

The report describes how half of all European citizens have heard of recent revelations about government agencies collecting personal data on a large scale for the purpose of national security (European Commission, 2015, p22). Of those that claimed to have heard about these revelations, a majority (46%) declare that this knowledge has a negative impact on their level of trust in how online personal data is handled, while 40% claim the revelations have no impact.

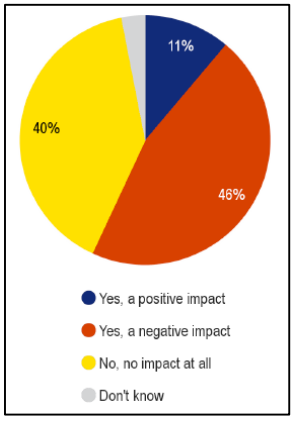


Figure 2: Impact of recent revelations on trust in personal data handling (European Commission, 2015, p.25)

2.1.2. Disclosure of Personal Data

Attitudes towards disclosing personal data

The level of discomfort about the disclosure of personal data amongst the 27,980 European Internet users was also evaluated. They found that a majority of respondents (57%) disagreed that providing personal information is not a big issue for them, whereas only 35% agreed with this statement. Furthermore, 52% disagreed that they do not mind providing personal information in return for free services online. Less then a third (29%) of the questioned respondents agreed with this statement (European Commission 2015, p.28). In the same report it is mentioned that 71% agree that providing personal information is an increasing part of modern life.

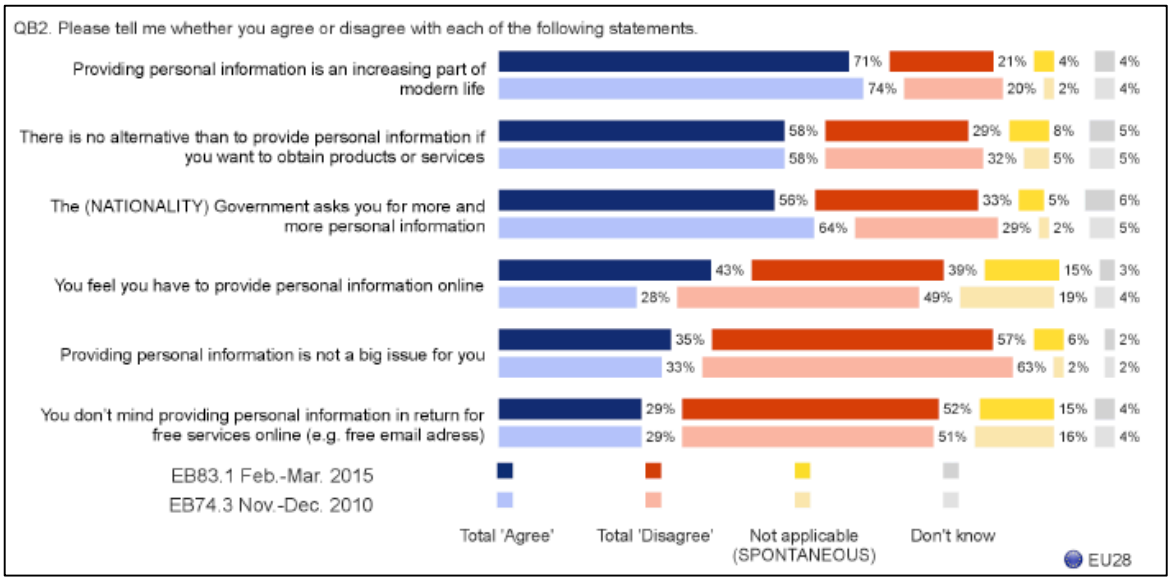


Figure 3: Attitudes towards disclosing personal data (European Commission, 2015, p.28)

2.1.3. Attitudes towards third party handling personal data

Personal data for tailored advertising

The survey also questioned the users' attitudes towards third party use of personal data and more particularly towards tailored advertising. This section is only relevant for participants who made use of the Internet, and was subsequently filled in by 21,707 participants. 53% of the respondents say they feel uncomfortable about the use of personal data by Internet companies for tailored advertisements, of which 17% feel very uncomfortable (European Commission, 2015, p.39).

Trust in authorities and companies on protecting personal data

The respondents were also questioned about their level of trust in various authorities and private companies to protect their personal information. More than one in two people trusted health and medical institutions (74%), national public authorities (66%), banks and financial institutions (56%) and European institutions (51%). Less trusting individuals were found with regard to the protection their personal data receives from shops and stores (40%), landline or

mobile phone companies and internet service providers (33%) and online businesses like search engines (24%) (European Commission, 2015, p.63).

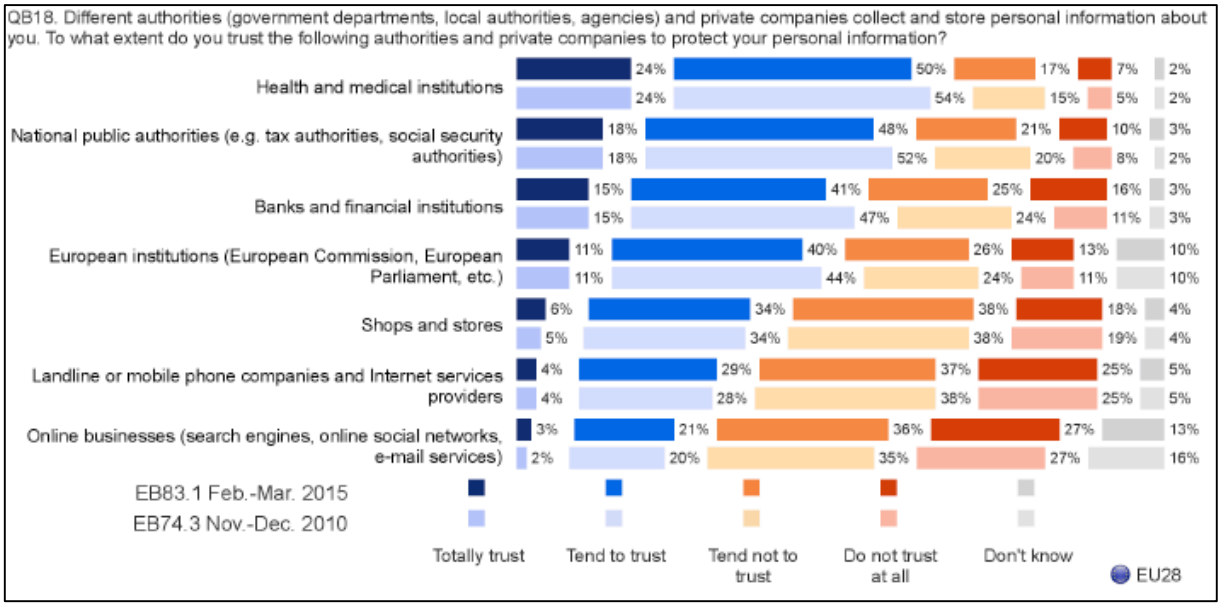


Figure 4: Trust in different institutions for protecting personal data (European Commission, 2015, p. 64)

Moreover, more than two-thirds of the respondents are concerned about their information being used for a different purpose than the one it was collected for (European Commission, 2015, p.68).

Responsibility for ensuring personal data is safely collected, stored and exchanged

The users were also asked who they think has the responsibility over safe collection, storage and exchange of personal information. Users were asked to provide two answers, for the first and second most important responsible parties (European Commission, 2015, p.104). The majority points to the individual him/herself as first responsible, followed by online companies; public authorities only come on the third place. The results can be found in detail in figure 5 below.

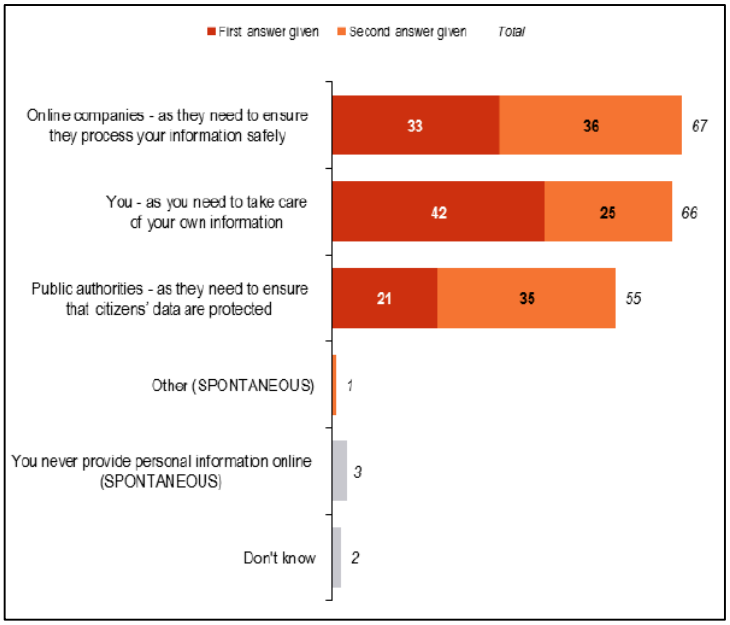


Figure 5: Responsibility over safe collection, storage and exchange of personal information online (European Commission, 2015, p. 105)

2.2. Online Information Disclosure

In the previous section, we provided a short overview of the most relevant results that are listed in the Eurobarometer report on data protection. This was essentially a short overview of the attitudes and beliefs of users towards personal data management and disclosure. In this section we look at a number of further issues with respect to online information disclosure. We first look at a classification / characterization of personal information. Subsequently, we discuss the fact that a minimum amount of information disclosure is always entailed when using OSNs. Finally, we investigate the value that personal data may have to the social media operators and other third parties.

2.2.1. Information types users share on social media

User data can be classified and categorized in different ways. For instance, in the annex of D3.8 the following types of user data were considered: registration data, incidental data, traffic data, interaction data and inferred data. We now consider an alternative characterization of data that considers two criteria: a) the meaning they provide to a data subject and b) the method of collection. Thus, we first propose to characterize data according to the meaning they provide on a data subject and we distinguish between **self-referential** and **relational data**. Secondly, we outline three methods of data collection: **volunteered**, **logged** and **inferred** data. Eventually, we will see that we can associate the data categories that were identified in the annex of D3.8 to different combinations along these two dimensions.

Self-referential data is information required to register or define something or someone on a social media platform. It refers to self-referential attributes that define the object or subject. Examples are name, username, surname, age, gender, type of object, etc. What is important is that all of this data refer to *only one entity* and define it as such.

The second form of data is **relational** information. This is information that ties two subjects or objects together. For example, if person X befriends person Z, than X and Z are friends. Relational data can be one- or bi-directional. In the case of friending, X and Z both have to agree on this relation. In case someone is following another person, this is one-directional.

Note that these two forms of data are not entirely separable. The fact that X is a friend of Z also means that X has at least one friend, which is self-referential information.

As mentioned, there are three data categories according to the way the data has been collected: volunteered, logged and inferred. They refer to the amount of agency of a user in the disclosure process.

With **volunteered** data, the user has the most control with regard to what information is disclosed; he or she can choose to give information or not and may even decide to lie. In all cases, the act of disclosure is entirely controlled by the user.

With **logged** data, the user becomes a passive object in the disclosure process. The data is exposed on behalf of the user by his or her device, browser or app. In other words, the data collection occurs automatically and without, or with very little effort from the user. This also means that it may be more difficult for users to understand what is shared with whom. For example, with cookies it is clear that something is shared with the creator of the cookie but it is unclear what this is exactly. If users are presented with a choice in this data collection

method, then this choice consists of stopping or resisting this collection. This is for example done by installing browser plugins that block the monitoring of web browsing behaviour.

Lastly, with **inferred** data the user has even less control. This is so because inferred data result from the analysis of volunteered or logged data, taking into account knowledge about the average behaviour of users similar to the data subject. Therefore, users only have (full or partial) control over the data that are used to infer new user attributes and have no control about the types of inferences carried out. Sometimes, users have a say in the inference process; this is usually limited to acting upon these inferences by correcting or deleting them. For example, Google and Facebook[1] let users see what their advertising preferences are. These ad preferences are indeed based on volunteered and logged data.

The following matrix lists specific examples of information disclosure along the two considered dimensions:

	Self-referential	Relational
Volunteered	I am X	X is friends with Z
Logged	X logs in on location Y	X visits profile Z
Inferred	X may be a resident of Y	X is probably interested in store Q in Y

Table 1: Description of matrix of types of personal data shared online

And if we apply this to the data types described in the annex of D3.8, we beget the following synthesis.

	Self-referential	Relational
Volunteered	Registration Incidental	Interaction data
Logged	Traffic data	Traffic data
Inferred	Inferred data	Inferred data

Table 2: Populated matrix of types of personal data shared online

[1] www.google.com/ads/preferences and https://www.facebook.com/ads/preferences/edit.

2.2.2. Obligatory passage points as encouragement process to share data

There is a minimum amount of information disclosure that results from participating in an OSN. We frame this situation as an obligatory passage point (OPP), an OPP is a situation where one or more actors are forced into particular behaviour because there is no other way to obtain a certain result (Latour, 1992). The concepts stems from actor-network theory (ANT). We consider four specific OPPs and examine the data disclosed at each of them: a) registration, b) usage, c) data inferred by the OSN and d) data obtained by examining user settings. In the following we look at each of these and also link the relevant information to the types of data that we previously identified.

We also mention that we researched and frame social media as OPP to connect and communicate with peers (Heyman & Pierson, forthcoming). This means that in order to easily communicate in a social way, it has become difficult not to use social media. And in the case of teenage users, it even seems impossible (Heyman, 2015).

1. **OSN Registration process as OPP**

The following table lists the personal data required during registration by a number of popular OSNs[2]:

	Facebook	LinkedIn	Twitter
Name	First and last	First and last	Username
Email	Yes	Yes	Yes
Birthdate	Yes	No	No
Gender	Yes	No	No

Table 3: Data needed to register on social media

The table refers only to the minimally required information to create a working account on these social media platforms. After registration, more information is asked from users, but this information is not mandatory. It does not have to be mandatory since it is information most users will agree on disclosing in order to make the service more relevant. For example, adding where you are working, etc. help in finding relevant friends or people to follow.

We can characterize this information as self-referential or relational. With self-referential information we refer to data categories such as workplace, living area, interests, etc. while relational information refers to the social and who one is connected to.

2. **OPP during social media usage**

After registration and if people start using their social platform accounts, their usage is logged regardless of if any information has been voluntary disclosed or not. This logged information is disclosed by the browser or device of the user. Secondly, if users do decide to volunteer information, in the form of user generated content (UGC), then this content is also

[2] This data was obtained by following the registration process of these platforms in December 2015.

very likely to also contain logged information. For example, if users of Twitter or Facebook do not opt-out from location sharing, then their location is stored through Geo-IP[3] and their posts will also contain information such as the location and time of day.

In the case a user connects with someone or something, he or she creates relational data, for example user X likes object Y. But this also adds meta-information of object Y to X in the following form: X is a person who expressed an interest in objects like Y.

The user cannot stop the OSN from logging and processing such information, therefore we also identify the automated logging that takes place during OSN usage as an OPP.

3. Inferred data

OSNs perform inferences about their users regardless of the user approving it or not. Therefore, having the OSN analyzing their users' data is an OPP. Facebook and Google do this extensively, for example, they allow seeing what they have inferred about a user in terms of advertising preferences. Figure 6 below shows the ad preferences pane of Facebook:

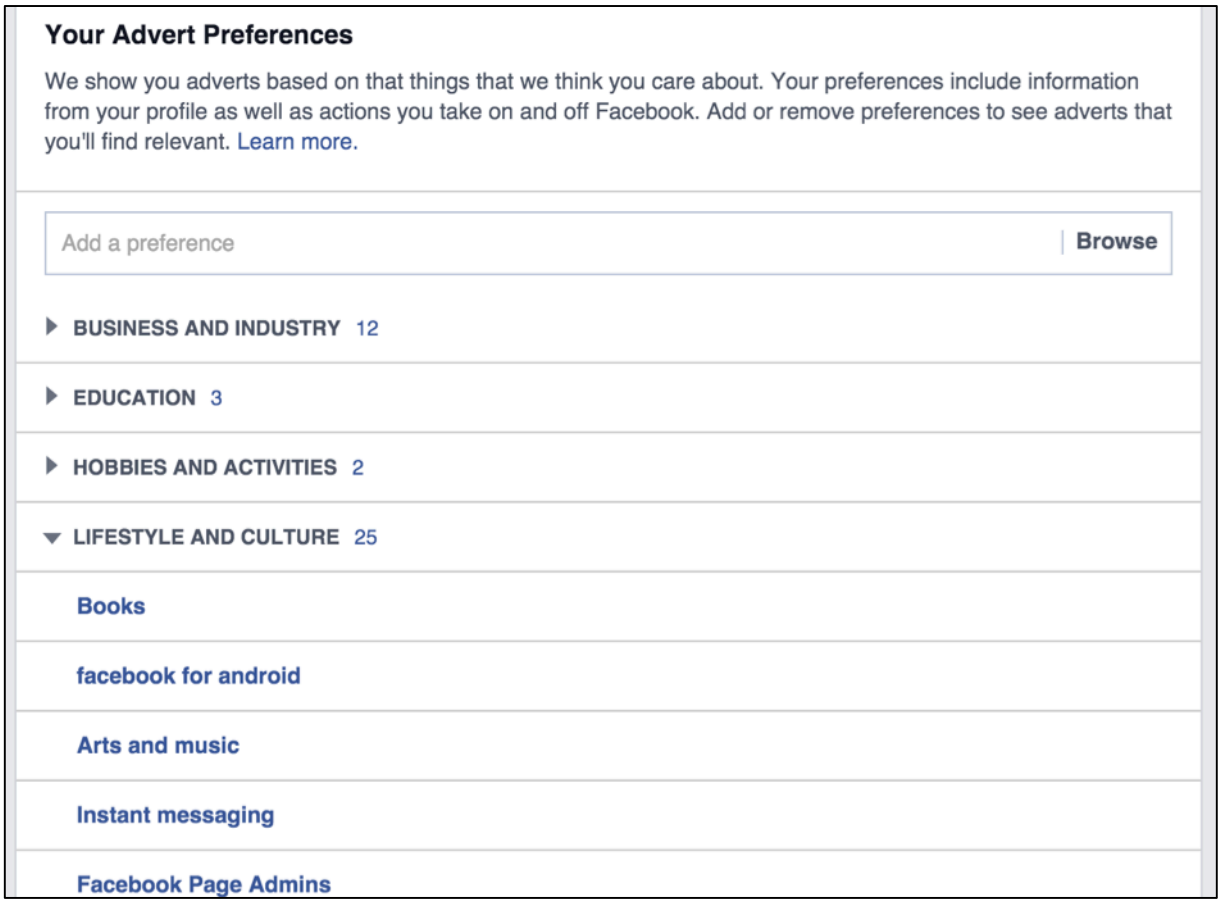


Figure 6: Example of Facebook's Ad Preference Pane

4. Default settings

Social media also steer disclosure of personal information through settings. Most often uninformed disclosure of personal information is encouraged. Non-disclosure is more difficult to achieve and for some information even downright impossible because there are no

[3] https://www.maxmind.com/en/locate-my-ip-address.

settings to change. Heyman et al., (2014) described this for social media platforms Facebook, LinkedIn and Twitter. For all of them:

- There are more options for interpersonal privacy than institutional privacy[4].
- If settings exist, then the default is as public as possible.
- It is impossible to opt-out of targeted advertising.
- Only informed users are able to opt-out of settings.

5. Conclusion

If we consider all these four points (registration, usage, inferred data and default settings) then we can see that users of social media have to volunteer a certain amount of data in order to render the service enjoyable. Moreover, each separate visit to social media consists of another OPP where data is logged. In conclusion, we can say that it is impossible to not share any data while using social media.

How does this relate to the EU Barometer results we cited in the introduction[5]? “Over seven out of ten people (71%) agree that providing personal information is an increasing part of modern life, slightly down from 74% who said this in 2010.” and “Just under six out of ten people (58%, no change) agree that there is no alternative other than to provide personal information if you want to obtain products or services" (European Commission, 2015, p. 28). Thus, we can argue that these respondents successfully recognized this minimum data disclosure as an OPP; it has become impossible not to share any personal data at all when using social media.

2.2.3. The value of personal data for social media companies

Now we examined how it is impossible for social media users to not disclose any data, we want to see why they are endorsed to reveal such an amount of data. The value of data for social media consists at least out of four dimensions. First, the data in the form of interpersonal communication and User Generated Content (UGC) renders social media into an OPP for this communication. Secondly, this data is used to personalise the flood of UGC so that it remains relevant content, which supports the usefulness of the first point raised. Third, this data is used to single out audiences for messages with promotional or editorial content and lastly, the data itself is valuable for other insights.

1. OPP for social communication

Many users of Facebook feel they would miss out certain information or social events because these are exclusively shared on Facebook. As a result, it has become difficult not to be on Facebook. This is illustrated in the quote below, where we asked respondents of 18 years old whether they would consider leaving Facebook:

[4] Institutional privacy refers to privacy settings to limit the flow of information towards organisations instead of particular people.

[5] It is important to note that personal information disclosure was asked in general and not with regard to social media.

Anna, 18: “I would only abandon Facebook if for example everyone would stop using Facebook. But leaving on my own, I can’t support the thought, because then you would lose a part of information you wouldn’t get otherwise.” (Heyman, 2015)

Bucher (2012) goes further and found that users are compelled to disclose a particular kind of UGC that is engaging for other users. Due to a feared threat of invisibility, users have internalised the rules of Facebook’s content aggregation algorithm. This has made users aware of the fact that they have to post engaging UGC on a regular basis if they wish to appear in their friends’ feeds.

2. Personalisation in a flood of UGC

Another issue with social media is the overload of UGC. As shown in the quote above, social media can only work as an OPP if they truly are an important source of platform exclusive information. In order to stay relevant for their users, social media use personal data to create a personalised selection of UGC for each user. Here algorithms require data to predict which content is relevant and which is not.

3. Attention scarcity

In this situation of information overload, attention becomes a scarce commodity and social media platforms capitalise this scarcity. Users, publishers and advertisers are able to promote their content if they pay. This means that their content is more likely to be shown to their target audience or friends than that of competing actors.

4. The data itself

The data itself has value. The data on social media also have value as a research object. Twitter sells access to its real time data and Facebook is known to use its data in controversial experiments (Kramer et al., 2014).

5. Conclusion

Much of the data users volunteer, is shared for the service social media provide. This service allows users to communicate with each other and to create UGC. But this is also an OPP to re-use the data for goals the user does not necessarily agree with, such as advertising and other big data applications. Since users have limited privacy management options due to missing and default privacy settings, all they can do is accept the secondary use of their data or leave the service.

2.3. Inference Mechanisms and Data Mining in USEMP

Inferred data was previously defined (see section 2.2.1 above) as data over which the subject has no control in the disclosure process. Based on previously logged or volunteered data, new attributes are inferred. In the current section we take a closer look at the mechanisms allowing this process.

2.3.1. Examples of inference possibilities on OSNs

Through social network sites, a vast and ever-increasing amount of data is made available. A lot of research is directed at discovering patterns in data from OSNs and if the way people interact with social media reveals something else about them. In this section we take a look at three interesting papers where a link was made between the use of social media and gender, distinguishing different social ties and social capital.

In a recent paper (Wang, Burke, & Kraut, 2013) it was investigated if users from different genders could be distinguished based on the topics that they are interested in. They found out that for the participants that were older than 25, women's posts are disproportionately more frequently about relationships and personal details, whereas men's posts are more likely to mention sports and abstract concepts such as politics and deep thoughts. For teens (age 13 to 17) Wang et al. found a higher similarity between genders based on the topics that they are interested in. They also found that men receive fewer comments from their network on their posts.

Another paper on Facebook data investigated whether it would be possible to recognize the significant other of a person based on the network structure (Backstrom & Kleinberg, 2014). It appeared that this was possible with a high accuracy based on a network measure called dispersion. This concept looks at the number of mutual friends of two persons (embeddedness) on the social network, together with the network structure on these mutual friends: "A link between two people has high dispersion when their mutual friends are not well connected to one another" (Backstrom & Kleinberg, 2014, p. 2).

(Burke, Marlow, & Lento, 2010) find that the greater OSN use is associated with increased social capital, defined here as the benefits made possible by the existence of a social structure, and reduced loneliness. In the paper they make a difference between two activities of OSN use: directed communication and consumption. Directed communication entails the interaction between two data subjects. Consumption is defined as the monitoring of all content that is not specifically targeted at the data subject (Burke et al., 2010, pp. 1–2). They found that bonding social capital (social capital between like-minded, homogeneous groups) is increased with the amount of direct communication. Surprisingly, they discovered as well that consumption leads to reduced bridging social capital. On the concept of loneliness they found the following: "People who feel a discrepancy between the social interactions they have and those that they desire tend to spend more time observing other people's interactions" (Burke et al., 2010, p.4).

These three examples indicate that by analyzing the use of social media sites, patterns can be discovered that might reveal more information about users than they may think they are disclosing.

2.3.2. Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) is the field of study that examines the process of derive knowledge out of databases (Fayyad, Piatetsky-Shapiro, & Smyth, 1996)(Heyman, De Wolf, & Pierson, 2014). The general goal of KDD is the "nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad et al., 1996), p.40-41; (Heyman et al., 2014). As described in these papers, KDD can be used for either verification or the discovery of new information. "In the first type, the data are used to verify existing hypotheses, and in the second type, the data are used to predict whether someone belongs to a certain profile (such as customer groups etc.)"(Heyman et al., 2014, p. 5). One of the purposes is then of course of a commercial nature, such as tailored marketing as the interests of the subject can be inferred and monitored.

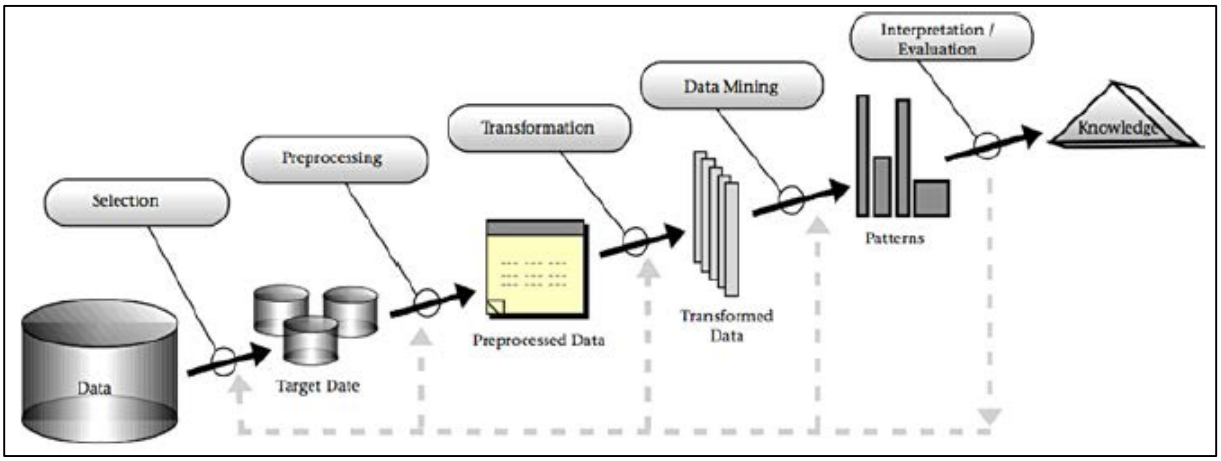


Figure 7: KDD Process (Ref. Fayyad et al. 1996)

2.3.3. Existing methods for inferring personal information

An overview of existing methods for inferring personal information based on OSN data together with their reliability can be found in D6.1, section 2.3. The next table summarizes these methods; it lists what is possible to infer through which method and based on what kind of input.

Method	Description	Input	Inferences	Dataset
Kosinski et al., 2013	SVD + Liner/Logistic regression	Likes	Demographics, Psychometrics, Habits, Preferences	Facebook
Schwartz et al., 2013	Differential Language Analysis (text) + PCA + SVM	Text of posts	Demographics, Psychometrics	Facebook
Backstrom & Leskovec, 2011	Supervised Random Walks	Friendship network, communication profile	Future friendship	Facebook
Backstrom & Kleinberg, 2014	Dispersion + Boosted Decision Trees	Friendship network, Demographics	Single/Married, Spouse	Facebook
Jernigan &	Friend attributes +	Friendship network,	Sexual Preferences	Facebook

Mistree, 2009	Logistic regression	Sexual Preferences of friends		
Zheleva & Getoor, 2009	Social network features + SVM	Friendship network, social features (membership to groups), partial labels	Country (Flickr), Gender, Political views (Facebook), dog breed (dogster)	Flickr, Facebook, dogster
Popescu & Grefenstette, 2010	Location Gazzetteer and Gender Vocabulary	Photo title and tags	Gender, Home location	Flickr
Rao et al., 2010	Text features + SVM	Text of tweets	Demographics, Political views	Twitter
Conover et al., 2011	Text-interaction features + SVM + label propagation	Text of tweets, retweet and mention network	Political views	Twitter
Pennacchiotti & Popescu, 2011	Text-social features + Gradient Boosted Decision Trees + label updating	Text of tweets, profile information, friends and replies network	Ethnicity, Political views, Starbucks fans	Twitter
Wagner et al., 2013	Text-social features + Random Forests	Text of tweets, profile bio, twitter lists	Profession, Personality features	Twitter

Table 4: Overview of personal information inferences approaches (Ref. D6.1)

As Goldsmith rightfully stated in 2012, “There are numerous tools and methods under development that claim to facilitate the extraction of specific classes of personal data from online sources, either directly or through correlation across a range of inputs” (Creese, Goldsmith, Nurse, & Phillips, 2012). From the disclosure dimensions distinguished in WP6 of the project, anything can be extracted in one way or another. What exists is only limited by the privacy awareness of the extracting agent and the profitability of the output. Therefore, it is interesting to see what the value of the different types of dimensions is. In the next section we will take a look at what the relevant information is for the advertising sector.

2.4. Valuable attributes for targeted advertising

In Heyman & Van Dijk (2013) the authors research what data is available as targetable categories for advertisers in Facebook's self-service advertising menu. Here we see that the following volunteered, self-referential data is directly targetable (if the user volunteered to disclose them):

- Location
- Age
- Gender
- Interests
- Sexual preference
- Relationship status
- Languages
- Education
- Workplace

The authors show that there are also relational criteria to target (Heyman & Van Dijk, 2013, p. 31). It is possible to target people with a certain relation to a certain object or subject. For example, it is possible to target people who have already liked a certain page, or to target people who did not do so yet. What is more, it is possible to target the friends of someone who already liked a certain page, to aid word-of-mouth promotion.

- Connections
    - Anyone
    - Only people connected to the promoted event
    - Only people not connected to the promoted event
    - Advanced connection targeting
- Friends of connections (target people who are connected to)

Secondly, all the objects users have liked become relational information as well, the authors point out that if a user Y has liked object X, that all the data pertaining to object X also become related to user Y. For example, if user Y liked a good cause event related to cancer, which is object X and if X is defined as an event that was related to the disease cancer Z; then for advertisers it is now possible to target Y as a person who expressed an interest in cancer Z.

Advertisers can also search for particular interests within the advertising interface. It is possible to segment users according to sensitive categories such as sexual orientation, political affiliation and medical information. These audience segments are defined as "People who have expressed an interest or like pages related to X" Where examples of X are a particular political party or for example cancer as shown in Figure 8.

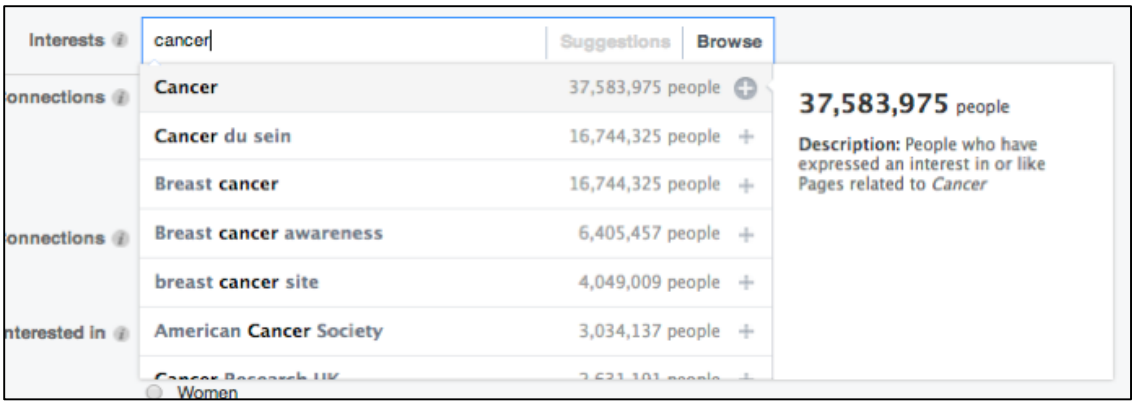


Figure 8: People who have expressed an interest or like page related to cancer (Ref. Heyman & Van Dijk, 2013, p.33)

Facebook states that this data certainly is not self-referential and does not reveal anything about the identity of their users: “We do not use sensitive personal data for advert targeting. Topics you choose for targeting your advert don't reflect the personal beliefs, characteristics or values of users. “But if people express an interest in multiple events with the same disease, for example, then it could be inferred that this person or someone in his or her vicinity probably has cancer.

Heyman and Van Dijk (2013) also refer to logged data in Facebook’s advertising categories: "For example, each browser automatically transmits which fonts are installed, what type of browser it is (Firefox, Chrome, Safari or Internet Explorer), the resolution of a computer screen, the operating system and so on. With this information Facebook can identify users of old computers for example.” They also show that advertisers are able to target users of specific iPhone or android type phones (see Figure 9 below).

Lastly, they also provide examples of inferred data through the categories “away from family” or “long distance relationship”. In both cases a combination of volunteered data (in a relationship), logged data (GeoIP to log the location) and the relational data with whom one has an amorous relation or family tie is used to see if locations are far enough from each other to define it as a long distance. It is impossible to sum up all the possibilities of inferred data because these options are limitless. Figure 9 is an overview of the categories that Facebook offered in 2011.

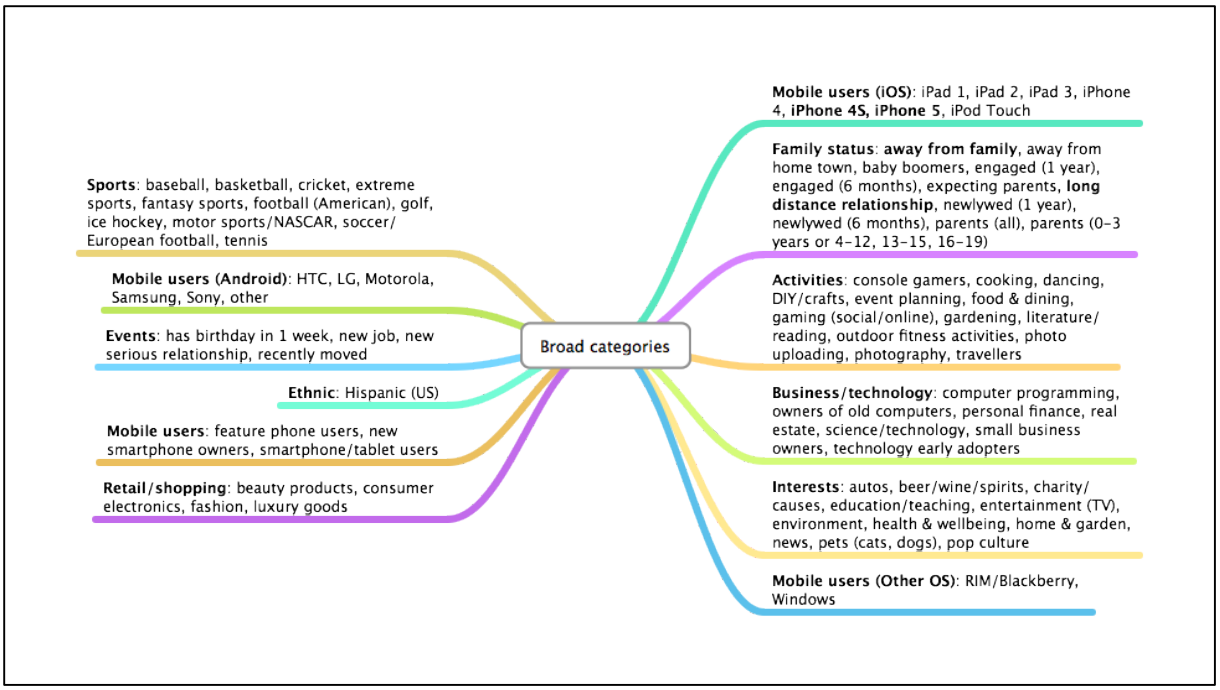


Figure 9: Facebook's Broad Categories (Ref. Heyman & Van Dijk, 2013, p.32)

3. Current Research in USEMP

Through means of the literature study presented in the previous chapter, an overview was created about users' attitudes towards data disclosure, how users are encouraged to reveal information online, which information can be inferred from online disclosures and how this creates value for the social media platforms and third parties. In our own research we build further upon the results of the previous section. In this chapter we first take a look at how the current research connects with the USEMP project as a whole and the DataBait tools in particular. Then we take a look at the set-up of the different research tracks before discussing the results in the next chapter.

3.1. USEMP Disclosure Scoring Framework

One of the main components of DataBait is the disclosure scoring framework. In the scope of the social research performed, we will not go into too much detail of the workings of this framework, an elaborate description of its workings can be found in both deliverables 6.1 and 6.4 of this project as well as in Petkos, Papadopoulos & Kompatsiaris (2015). We will only provide a quick overview in the next paragraph. We are more interested in how our research might have an impact on this framework.

At the core of the disclosure scoring framework, there is a set of eight disclosure dimensions. These are categories of personal information that people may disclose through their OSN behavior. The eight disclosure dimensions are: demographics, psychological traits, sexual profile, political attitude, religious beliefs, health factors & condition, location and consumer profile. Each dimension is made out of a set of attributes (e.g.: for the 'demographics' this includes gender and age). The attributes were chosen because they may be considered sensitive from a user perspective or legal perspective or when they hold value for marketing companies. Each attribute can take a number of values. An overview of the disclosure dimensions and corresponding attributes can be found in table 2. Eventually, this organization creates a hierarchic structure of dimensions, attributes and values, as shown in Figure 10. Please note that at each node at each level of the hierarchy is associated with a number of scores that express different aspects of information disclosure at the OSN. For instance, there is a score that expresses the sensitivity of some dimension, attribute or value and there is an overall disclosure score that expresses the overall risk associated with the disclosure of some type of information. Importantly, the scores are associated with actual data shared by the user on the OSN and these associations may result from the fact that the user has explicitly stated that some value holds for him / her (volunteered data) or may be the result of some inference process. As mentioned, for more details please refer to the aforementioned documents.

We applied these dimensions and attributes in both our quantitative and qualitative research when questioning the participants about their perceived sensitivity and willingness to share different types of information.

Disclosure Dimension	**Attributes**
Demographics	Age
	Gender
	Ethnicity

	Literacy level
	Occupation
	Income level
	Family status
Psychological Traits	Emotional stability
	Agreeableness
	Extraversion
	Conscientiousness
	Openness
Sexual Profile	Relationship status
	Preference
	Multiple partners
	Habits
Political Attitude	Political parties
	Politicians
	Stance in issues
Religious Beliefs	Supported religion
Health Factors and Condition	Smoking behaviour
	Drinking behaviour
	Drug use
	Chronic diseases
	Mediating factors (e.g. exercising)
	Medical history
Location	Home address
	Work address
	Favourite places
	Visited places
Consumer profile	Preferred products
	Brand attitude
	Hobbies
	Devices

Table 5: Overview of the disclosure dimensions and corresponding attributes

Now, how does the current research feed back in to the USEMP disclosure scoring framework? The research presented below will account for valuable insights for the generation of the sensitivity scores. As part of the quantitative and qualitative research, we questioned the perceived sensitiveness of the different disclosure dimensions. This will feed back into the backend of the DataBait tool as the sensitivity scores, which influence the eventual dimension disclosure scores.

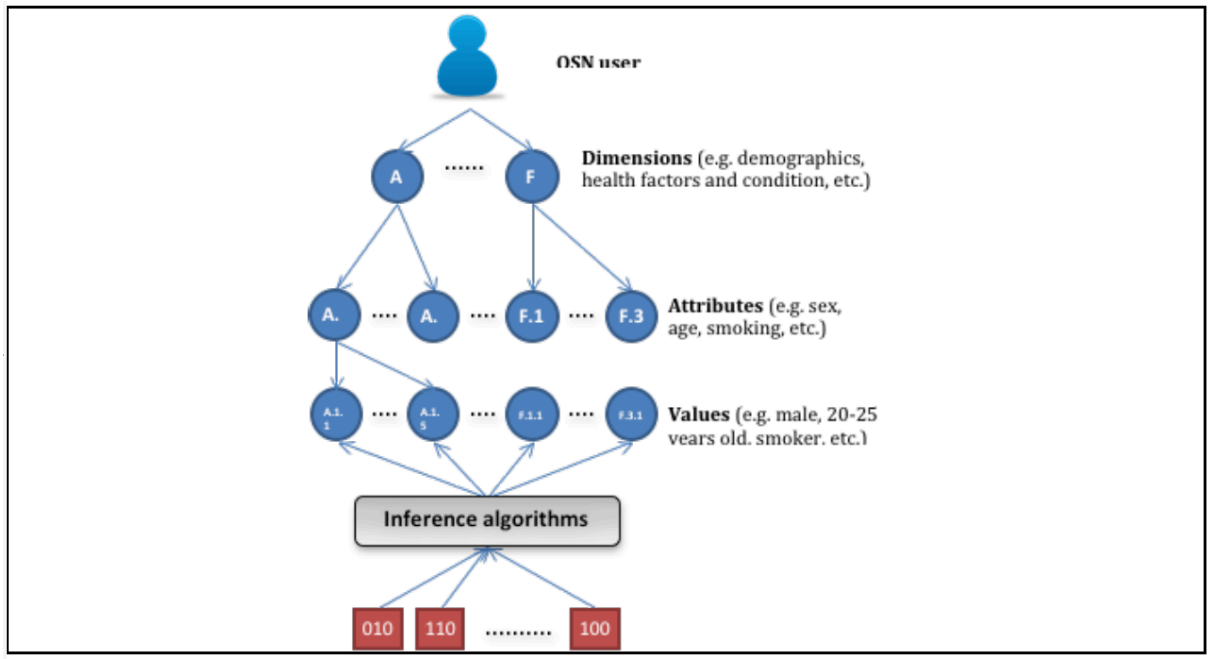


Figure 10: USEMP Disclosure Scoring Framework (Ref. Popescu et al., 2015)

4. Quantitative research Track

As part of the early pilots of the project, the OSN data from a number of users were collected, as well as their responses to a survey that questioned them about their personal details and their attitude towards disclosure of their personal information. In this chapter we look closely and in a quantitative manner at the data related to the users' attitude towards information disclosure.

4.1. Quantitative research track: Set-up and workflow

The aim of this research track was to receive access to 200 private Facebook accounts and its data. In order to do this iMinds set up a workflow in close collaboration with the other partners of the consortium. The recruitment was done with support of the Living Lab institutions of the two social science partners: iMinds Living Labs and Botnia. The research track started at the end of April 2015 and continued throughout the month of May. The consortium made use of a beta-version of the DataBait tool to get access to the Facebook profiles of the participants. For this reason, this quantitative research track was named the pre-pre-pilot, due to the maturity of the tool at that ime. Because Facebook did not allow DataBait to be a public application yet, the participants needed to be added manually as test users to the application.

The Living Labs institutions sent the invitations to participate on April 21st, 2016. Through the invitations they were redirected to a web form where they were asked to fill in their Facebook IDs. As a second step, iMinds added them as test users and resent the participants a second email that they could now register on DataBait and link their DataBait and Facebook profiles, in this way granting us access to their data. As a final step, the participants had to fill in a survey, where they were asked questions relating to all eight Disclosure Dimensions. In this way we could link the Facebook data with the data they claimed to be true in the survey for the training of the inference algorithms. So in order for this research track to be successful, we needed to gain access to the respondents' Facebook data and have them fill in the survey. The complete flow is also visualized in Figure 11.

Figure 11: Workflow of the quantitative research track

Due to the extensive tasks we asked our participants to carry out, we got a significant amount of dropout over the course of the process. Of the 255 respondents who agreed to partake in our research, only 182 (or only 71%) respondents fully completed the task. A lot of dropout occurred between step 3 and 4. Here the respondents had to wait before they were being added as a test user and could continue the process. Another moment of significant dropout was in the survey, where they were questioned about their psychological traits, probably due to the large amount of questions we had to ask to have a proper scale for this

feature (The Big Five Inventory consisting of 42 items, see e.g. (Denissen, Geenen, van Aken, Gosling, & Potter, 2008a)). Please see the annex for an overview of the invitations (see 8.1) and survey (see 8.2).

4.2. Quantitative research track: Results[6]

In the survey we asked our participants for information about the eight privacy dimensions: demographics, psychological traits, sexual profile, political attitude, religious beliefs, health factors & condition, location and consumer profile. In order to get more fine-grained information and to present the user with more digestible blocks of questions, demographics was further split up in: basic demographic information and professional & financial information, and sexual profile was split up in relationship information and sexual information. Furthermore, psychological traits was renamed personality traits. Eventually, the participants were presented with 10 blocks of questions as listed below. The full questionnaire can be found in the annex of this deliverable (see 8.2).

Information asked in the survey
Basic demographic information
Professional and financial information
Relationship information
Religion
Personality traits
Political attitude
Health and condition
Location information and holiday preferences
Brand preferences and interests

Table 6: Information in the survey

After each block of questions we asked them the following three questions:

1. How **sensitive** do you find the information you had to reveal about your (**information type**) in the previous section?

   → (Likert scale: 1=not sensitive at all, 7=very sensitive)

2. How **important** is it for you that this type of information about you remains private?

   → (Likert scale: 1=not important, 7=very important)

3. Do you think the information on your Facebook profile **reveals** this (**information type**)? Either because you yourself have put it online, or it could be inferred from a combination of posts.
   → (Yes / No / No answer)

[6] A first overview and analysis of the survey data can be found in (Petkos, Papadopoulos, & Kompatsiaris, 2015), whereas a first examination of the use of the data for training a set of inference modules can be found in D6.4.

The answer to these questions then gives us an insight in the perceived sensitivity, perceived disclosure and how classified the different information types should remain.

4.2.1. Perceived Sensitivity of the Disclosure Dimensions

We questioned the perceived sensitivity of the ten disclosure dimensions on a 7-point Likert scale, with 1=not sensitive at all, 7=very sensitive. In Figure 12 below, an overview of the averages is presented. Health information, professional information and personality traits are among the types of information that our population perceived as the most sensitive on average. Demographics, religious and sexual information received the lowest average scores. In our survey, professional information questioned both the employment status and income level of users. It is suspected that the latter explains the high perceived sensitivity.

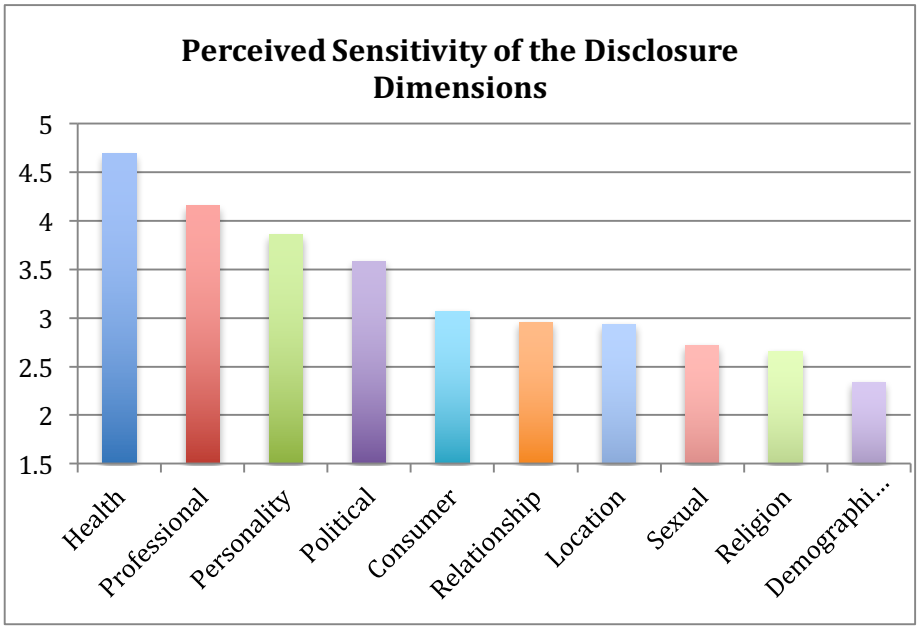


Figure 12: Perceived Sensitivity of the Disclosure Dimensions, Likert scale 1-7 averages, n=163

4.2.2. Perceived necessity of keeping the disclosure dimensions private

The second feedback question was about how important it was to the users that the information related to some dimension remains private. We again asked this question on a Likert scale from 1=Not important to 7=very important. The outcome of this perceived necessity of keeping the different disclosure dimensions private is in line with how sensitive the information was perceived. With health information, professional information, personality traits and political information being perceived as the information types that our participants preferred more to keep private.

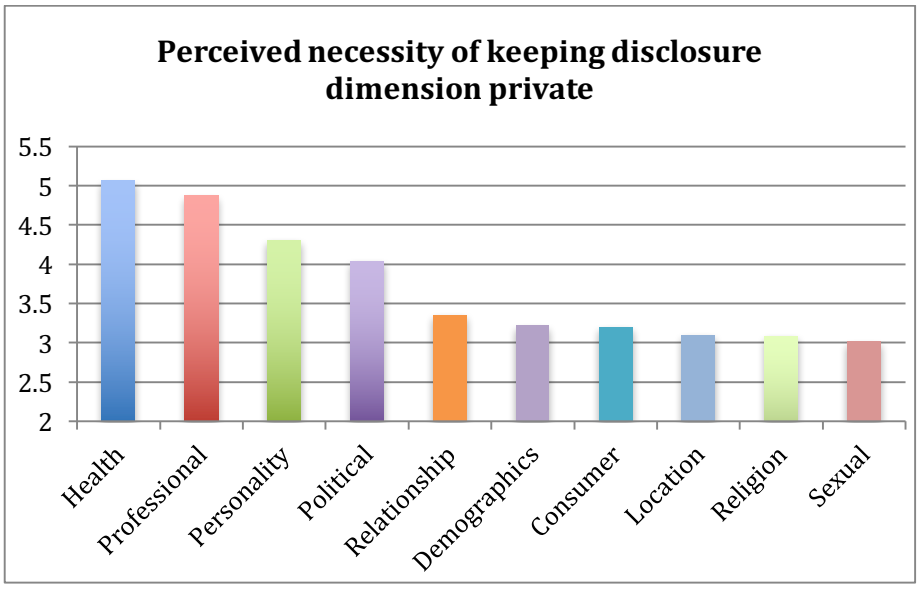


Figure 13: Perceived necessity of keeping disclosure dimension private, Likert Scale 1-7, averages, n=164

4.2.3. Perceived Predictability of Disclosure Dimension

Finally, we also asked our participants to answer the question: "Do you think the information on your Facebook profile reveals your (information type / dimension)? Either because you yourself have put it online, or it could be inferred from a combination of posts." This was a plain yes or no question, with a third option of not answering. Figure 14 below presents the percentages of the population that answered the question that thought the disclosure dimensions *could* be found on their Facebook profiles. Demographics (88%), Location (87%) information and Relationship information (85%) were the disclosure dimensions the majority of our population thought is revealed through their Facebook profiles. On the other side of the spectrum we found that least amount of people thought Religion (49%), Professional (47%) and health information (37%) were disclosed. In deliverable 6.4, a comparison was made between perceived predictability and actual predictability resulting from the inference experiments done in USEMP. For sake of completeness this table is also included below, for more information please see D6.4: USEMP disclosure scoring framework and disclosure setting framework – v2.

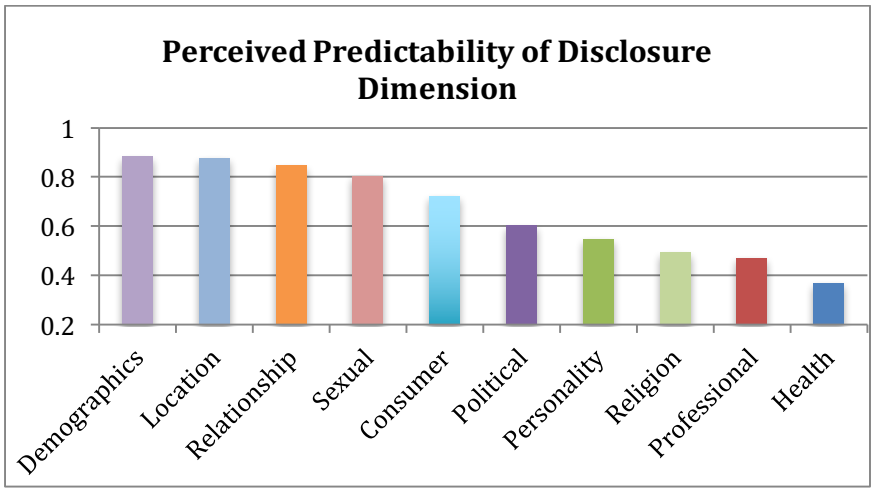


Figure 14: Perceived Predictability of Disclosure Dimensions, n=161

Ranking	Perceived predictability of disclosure	Actual predictability of dimension
1	Demographics	Demographics
2	Location	Political views (+4)
3	Relationship status and living condition	Sexual orientation
4	Sexual orientation	Employment status and income (+5)
5	Consumer profile	Consumer profile
6	Political views	Relationship status and living condition
7	Personality traits	Religious views (+1)
8	Religious views	Health status (+1)
9	Employment status and income	Personality traits
10	Health status	

Table 7. Ranking of perceptions of users about the disclosure of each dimension and ranking of predictability of each dimension according to the experiments presented in D6.4 (See D6.4)

5. Qualitative research Track

Besides the quantitative part, described in the previous chapter, which provided us with descriptive information about users attitudes towards the sharing of different types of personal information as categorized by the disclosure dimensions and how this compares to the actual predictability of those dimensions, we made use of qualitative interviews to gather more in depth information. Initially, 14 interviews were conducted to receive a deeper understanding of users' willingness to share different types of information (referring to the defined disclosure dimensions), furthermore we took a look at users' attitudes towards different types of information collection (volunteered, logged and inferred information) as well as at their attitudes towards five different institutional actors reasoning on their data (insurance company, government, advertising agency, (future) employer, academic organisation). Due to the fact that we wanted to have personal evaluations of all the above, we opted for individual interviews instead of a focus group of group interviews. The 14 interviews were later supplemented with 7 more for one specific task, as we will explain in the next section.

5.1. Interviews: Design

14 interviews were conducted in Dutch between 26 November 2014 and 19 December 2014 in the region of Flanders, Belgium. We opted for a selection of participants with maximum variance in age and gender. In order to be considered eligible, the participants needed to have a Facebook account. In September 2015, we conducted another round of interviews in light of a usability research for WP 8 of this project. When we had sufficient time, we asked the participants to perform one of the tasks of the interview again (the card sorts, see below). In Table 8 we see an overview of the participants, the 7 participants who performed the card sorting in the second round of interviews are marked with a star.

Pseudonym	Age	Professional Situation	Frequency Facebook use
Adam (m)	41	Employed	Monthly
Mélanie (f)	22	Student	Several times a day
Annie (f)	34	Employed	Daily
Mac (m)	24	Employed	Daily
Kendrick (m)	24	Student	Daily
Roger (m)	35	Employed	Daily
Kevin (m)	68	Retired	Daily
Joel (m)	24	Employed	Daily
Nina (f)	22	Looking for first job	Several times a day
James (m)	25	Student	Several times a day
Valerie (f)	24	Student	Several times a day
Tim (m)	22	Student	Daily
Louisa (f)	35	Employed	Several times a day
Bridget (f)	22	Student	Several times a day
Neil* (m)	37	Employed	Several times a day
Sarah* (f)	27	Employed	Weekly
Courtney* (f)	38	Employed	Several times a day
Paul* (m)	32	Employed	Daily

Bob* (m)	20	Student	Several times a day
Joni* (f)	23	Employed	Several times a day
Rick* (m)	53	Employed	Weekly

Table 8: Overview of participants (Pseudonimised)

For the recruitment of respondents, iMinds relied on the expertise of iMinds' living labs. As an extra incentive for participation all participants received a voucher of €25 for a Belgian retail chain for books and multimedia (FNAC). All respondents signed an informed consent at the start of each interview, in which the research was contextualized and it was stated that all information would be pseudonomised and handled with great care.

The total of 14 participants, which consisted of 8 male and 6 female respondents, were recruited in the region of Flanders. With an exception of 4 sessions, the interviews were carried out in the iMinds research centre in Ghent, 2 were hosted in the iMinds research centre in Brussels and two in a public location closer to the participants' home (respectively in Aalst and Leuven. The interviews lasted on average two hours. Table summarizes the participants' demographics. 13 sessions were tape-recorded and subsequently transcribed. One participant preferred not to be recorded; here notes were taken during and immediately after the session. To ensure anonymity, all respondents received pseudonyms in the transcriptions, which will also be used throughout the deliverable.

To leave ample time for discussion, we opted for semi-structured interviews, revolving around three main tasks. A script for this was prepared with a number of follow-up questions should the conversation stumble. The interviews were structured as follows:

1. Short introduction: explaining the voice recording of the session and the general outline of the interview (15').
2. Q-sort exercise: gather information on the respondents' willingness to share different types of information online in relation to the disclosure dimensions (30').
3. Bowls away exercise: gather information on the respondents' attitudes towards different institutional actors reasoning on their data (20').
4. Information deletion exercise: gather information on the respondents' attitudes towards different types of data collection and different institutional actors reasoning on their data (20').
5. Asking for possible additions and comments (10').

In the next sections, we will report the first results of each of the subsections of the interviews. The interviews were transcribed ad verbatim from the audio recordings. For the analysis we took a look at the different exercises to come with overarching results for attitudes towards different types of institutional actors reasoning on personal data, the attitudes towards different types of data collection. In the last section we take a look at the willingness to share different types of personal information.

5.2. Setup of the exercises

5.2.1. Bowls Away

The data gathered at the interviews stems from three different exercises. In the first one, referred to as bowls away, we asked our participants to divide nine types of personal information (connected with the eight disclosure dimensions, splitting op demographic information into socio - demographic information and financial information) and eight types of data input into three bowls. The main goal of this exercise was to get insights into the attitudes of our respondents towards sharing different types of information with different institutional actors. Each bowl accords to the participants' willingness to share the piece of information with an institutional actor: ok to share with the institutional actor – neutral - not ok to share with the institutional actor. We repeated the exercise for each institutional actor (advertising agency, government, (future) employer, insurance company and academic organisation). We quantified the data that was put in the bowls to get a first overview over the responses; this was then supplemented with the attitudes that the respondents were probed to say out loud while the exercise was ongoing.

Disclosure Dimensions	Data Input
Socio-demographic information	Status updates
Financial information	Google Search
Psychological traits	Pictures
Sexual information	Links
Political stance	Events
Religious beliefs	Private Facebook messages
Medical information	Used hardware and its location (observed)
Location information	Likes
Consumer profile	

Table 9: The different types of information for the bowls away exercise

5.2.2. Information deletion exercise

In the second exercise, named the information deletion exercise, we presented the different participants with a fictitious persona, Nora, who shares different types of information, again relating to the disclosure dimensions. Through this exercise we again received insights into the attitude towards information sharing with different institutional actors as well as the attitude towards different types of information collection (volunteered, observed and inferred information). Our participants could strikethrough the information that they didn't find suitable to share with each of the five institutional actors. An example of this exercise can be found in the annex (see 8.3). The exercises should be seen as a probe for the participants to get them thinking about the topic. They were urged to talk aloud during the performance.

5.2.3. Card sorting exercise

Thirdly, we took a look at the willingness to share towards different types of information. Each participant got 36 cards with different types of information, which they had to place on a grid. This exercise made it possible to get a first insight in the willingness to share of different

types of personal information. Later on we can perform a Q analysis on the cards so we can find out if it's possible to distinguish different kind of groups with different sharing beliefs and what variables define these social groups.

5.3. Qualitative research track - results

In this section of the deliverable we will present the results of the qualitative research track. We will subsequently take a look at our respondents' attitudes towards different types of institutional actors reasoning on personal information, their attitudes towards different types of information collection (volunteered, observed, inferred) and finally a first glance at the willingness to share different types of information in an online context.

5.3.1. Attitudes towards different types of institutional actors reasoning on personal information

First we will take a look at the attitudes of our 14 participants towards different types of institutional actors reasoning on their personal information. We made a distinction between five institutional actors for whom access to personal data of the general public is beneficial: the (future) employer, the insurance company, the government, the academic research organization and finally the advertising agency. After the interviews were conducted, we searched for the overarching and recurring ideas that our participants had towards the different actors, which we will present below together with some key quotes that underwrite the statements.

To get a first overview, we quantified the way our 14 respondents ordered the 17 pieces of information in the 'bowls away' exercise for each individual institutional actor. We gave a score to each information piece according to its willingness to share:

- Fine to share: 1
- Neutral: 2
- Not fine to share: 3

Later we took the sum to see if our small population brought already some different attitudes forth. It must be stated clearly here that this was done only to get a first generic overview. Due to the small number of respondents, the qualitative data is of more importance. In this way we could find out for each participant, which actor they were least willing to share information with.

10 of the 14 respondents were willing to share the smallest amount of information with the insurance company. Similarly 10 of the 14 respondents were willing to share the most amount of information with the academic research organization. The complete ranking is the following: they would want to share least with the insurance company, followed by the advertising agency, future employer, government and finally the academic research organization. In the next couple of sections we will take a look at each actor individually and see which attitudes surfaced and how our respondents motivated these.

A. (Future) employer

We presented our respondents with pieces of information that are available online for a fictitious person named Nora. We then asked them if they would agree that an employer or a future employer would take a look at the information and why?

The respondents sometimes consented with a future employer looking at online information about an applicant, as that it can help to make a better judgment if the person will fit in the team.

> *Roger (m, 37): "They want to hire a person, not a profile. In this way they know what type of person they have seating in front of them. You know immediately a little bit more about the person that you're going to have to work with. You can make an estimation if the other colleagues will like him."*

Or

> *Valerie (f, 24): "I think they all do it, I don't mind so much, it's a superficial impression. It's something that you decide to put online. Maybe they know what type of person the applicant is, and if they will fit in the company or not. I think it's fair that it's taken into account, because it's also for your own good. If you will belong in the company, if you will feel good in the team or not."*

But most of the respondents do agree that the (future) employer can only look at the online available information if it is **relevant** information, such as the studies, the language skills, location etc.

> *Kevin (m, 68): "A lot of the information is not relevant for the employer. They have to find it out for themselves when she's applying or when she works there for a couple of years. But not before they even met.*

And a lot of them also fear **discrimination** of employers looking at all the available online information.

> *Valerie (f, 24): "They definitely don't have to look at the political preferences. If you don't agree on that fact you're going to see the person in a different light. Religion shouldn't be a problem, but I'm not going to allow it, because it is sometimes also a reason for discrimination."*

Our respondent also made a difference between **the sources** where the information comes from.

> *Nina (f, 22): "I agree with them looking at my LinkedIn-profile, not my Facebook profile. LinkedIn is for this exact purpose and you can see that they visited your profile. I like it that I then know if they are interested.*

B. Insurance company

As mentioned before, of the five institutional actors the insurance company was the one with whom our respondents' willingness to share information appeared to be the lowest. The most reoccurring reason for this is that in this case the use of information can become very **personal** (instead of on an aggregated level) and it can be **used against them**.

> *Nina (f, 22): "I think there are a lot of people who have their pictures public. But if she's on there smoking: I don't think that's good for your life insurance. That wouldn't be smart."*
>
> *Adam (m, 41): "I think that you have to be careful with your psychological traits and medical information. Give away as least as possible, only the essential things and as little possible that might be in my disadvantage"*
>
> *Roger (m, 37): "I personally don't think they have the right to know a lot of things. They can know her name, address and that's it. The rest is none of their business. Especially because they are going to draw conclusions on the information and then you're going to have to pay a lot of money based on some conclusions that are not a fact."*

As this last quote already underwrites, our respondents only agree to share information with an insurance company if it's **relevant** for them to know, like a cell phone number or if it's **not harmful** or **beneficial** for the data subject included.

> *Mac (m, 24): "If it's purely anonymous, I wouldn't be bothered by it. If the insurance company would send a new police proposition to everyone with an abnormal heart rate, I wouldn't mind it if I was already a customer there."*

One of our respondents also noted that if an insurer addresses health problems that you never told him that the practice of looking things up online will be counterproductive

> *Nina (f, 22): "If you have a talk with your insurer, and this guy tells you that you aren't doing well with your health, you're going to start asking questions how he knows this. You will not have a lot of trust in this company, because you know he's brownnosing in your private life and sees that you're smoking. It will backfire."*

C. Government

With regards to a government institute, our respondents were in general a little bit more willing to share information (in comparison with the insurance company). The reason given is however not so much a positive one. They feel as if the government knows a lot of the information already in one way or another so they don't mind it so much.

> *Melanie (f, 22): "The government .. maybe a better question is: 'what don't they already know?' "*
>
> *James (m, 25): "You have to give them a lot of information anyway, so they'll know it. The things they already know I don't mind so much".*

One person was more positive and thinks the government can use this to govern a society better, or to create a plan for better prevention of crime and diseases. But the respondents also feared that it would run out of control the way it's developing now.

> *Annie (f, 34): "Eventually it might be possible to link shopping cards with health insurance. The shopping card will say: 'yeah, he always buys junk food and never fruit or vegetables, he's a potential threat and we must make him pay more for his health insurance.' "*

Roger (m, 35): "That's really not relevant for them. Religion and all the other things.. Looking at my pictures? No! That's almost like the Stasi, what are we doing then? No, they do not need to know that. Maybe the language she speaks, but that's it."

D. Academic Research Organisation

In general, our respondents were most willing to share with an academic research organisation. We have to of course take into account that some wanted to answer in a 'socially acceptable' way, being in the presence of a social scientist. The respondents that are willing to participate in a social science research may also be biased towards wanting to share their information for the purpose of science.

Notwithstanding these remarks the reasons they gave for their higher willingness to share with an academic research organisation was the **trust** they have in research for keeping the conclusions **nuanced.**

Tim (m, 22): "With research you know that they will not investigate the data in a vacuum"

Mac (m, 24): "Integrity, what they will do with all my data, I think they will use with confidentiality. There are enough smart people in a research organisation to know that they don't have to make to drastic inferences."

Another reason that was given is that the respondents believe that the data will be used in an **aggregated** way and that it does not become so personal.

Bridget (f, 22): "That will never be looked on an individual level, they will throw it together to perform some analyses on it. That's why I don't mind so much. […] They will not use it to approach you, and it will not be used individually."

Valerie (f, 24): "I know that they won't relate the data with your personal life. It's more about this amount of people to this, this amount of people do that, it's not so personal."

Some respondents also made the nuance that they feel that academic research organisation can only do this if the information is **relevant** for the specific research.

Tim (m, 22): *"It depends on the kind of research. I also don't think that I would give all the data for one survey for instance. I'd prefer to do it in pieces: for example, when there's election and I take part in a survey about politics then I would give away my political preference. If there's a survey about health, then I would say that I smoke."*

Roger (m, 35): "I would give all the data depending on the research at hand. For example in this interview I would not tell about my medical information, but for the testing of medicines they sometimes ask for blood samples and they ask medical questions, then I would give all the information they need. But I would never come tell that here."

E. Advertising Agency

For the advertising agency, we could distinguish two bigger groups. There were some respondents who would not share too much information with advertising agencies, they often mentioned how they see advertising as too **intrusive**.

> *Kevin (m, 68): "(If you go to the website of Paul Simon) And then they will advertise records of Paul Simon of from similar artists and then you have Amazon and other sites coming after you. There's hundreds of them".*
>
> *Bridget (f, 22): "I think it's too easy to be influenced by advertising. If they… I spend a lot of time on my computer, so if they know everything what I do, they know who I am. […] I become too vulnerable in this way if they know all this information. So I don't want to share anything with them."*

Those who do want to share information with advertising agencies see **benefits** in sharing and in receiving **tailored advertising**.

> *Roger (m, 35): "Some websites still do it, advertising of the lamest online games. I have never played them and they provide very ugly advertising on every inch of websites. Then I prefer that they know that I went to Coolblue, that I searched for an item and that they advertise that product again on another site."*
>
> *Tim (m, 22): "Actually I wouldn't mind, if they would provide me with good advertising. I would give every piece of data. That I smoke? Promotion for cigarettes. My address? So they can send me the advertising. Not my cellphone, they don't have to call me. That I have a new car? Then they can send me information about new accessories for the car. […] Medication for a lack of sleep, a booklet about health and heart problems."*

5.3.2. Attitudes towards different types of information collection

In our 'Information Deletion' exercise we also presented our respondents with the differences in data collection (volunteered, observed and inferred). We explained them shortly how this works and tried to probe them for their attitudes.

In general we found that our respondents had less of a problem with the use of information if it was **voluntarily** and **consciously** uploaded on the Internet, and that the fault is more with the data subject.

With regard to a future employer Kendrick for instance said the following:

> *Kendrick (m, 24): "If you don't want it, don't put it online. […] Nobody forces you to put something on Facebook. You may not always think about it all the time, but you do upload the information yourself.*

James gave a similar response:

> *James (m, 25): "I do think that when you put something public on a profile it's similar to standing on a square and start yelling. I don't find it ok that people think that they can post something online and then expect not to get judged for it. For instance if they insult their employer on Facebook, than that's a stupid mistake and their own fault. People should not think that they are protected.*

With regard to **observed information**, our respondents become more critical. Especially when talking about Google search or the websites one visits. Since people don't voluntarily

put it online and don't always think about the traces. Also they feel that there is also **no proper alternative** than browsing online if you want to look for something

> *Valerie (f, 24): "Google search, you do not even know yourself all the places you went to and what others might see. It is very unclear what people give away online. […] Otherwise you have to start looking for another way of searching for information. The Internet is made to make it easier for us and not to track us everywhere."*
>
> *Kendrick (m, 24): "I do not like to share this observed information. You don't share it directly, you do this in an unconscious way"*
>
> *James (m, 25): "For me it's off limits, that's really intruding and you can not do anything about it. It's intruding in what people do online and their history, and that's not ok. The volunteered information is different; it's neither illegal nor unethical to see what others have shared. […] Nobody may snoop around in that, in the history and the websites they visit, it's really like an authoritarian state then.*

Most of the respondents did not at all agree with the process of **inferring** other data that one has not voluntarily given online. The reasons given were that you do not have control and that it might be plain wrong information. They also noted that you suddenly can be categorized without knowing why and that others might find out information about yourself, that you yourself didn't even know.

> *Valerie (f, 24): "I find that inferred information is the most dangerous, because it's not by definition right information. If you fill something in yourself, than you know that it's a fact"*
>
> *Annie (f, 34): "Imagine that you get advertising that is very tailored, I would find that very creepy, because you will start thinking: how do they know all this? […] Especially when talking about medical stuff."*
>
> *Mac (m, 24): "I find it very dangerous. With everything that you do online, all your likes, stars, what you buy, other things that you reveal companies can make conclusions and you have no idea what it's based on. By giving a star too much you may find yourself categorised in another group. In general I think it's kind of dangerous"*
>
> *Bridget (f, 22): "It's very crazy, because you might not even know things about you yourself that others can infer".*

5.3.3. Willingness to share different types of information in an online context

Our central goal of the 'Bowls Away' exercise was to gather the attitudes of our respondents towards the willingness to share different pieces of information with a set of institutional actors, as a by-product however we also gathered some first insights into the willingness to share information in general. The 17 pieces of information questioned were nine types referring to the eight disclosure dimensions (we split up demographic information into socio-demographic and financial information) and eight types of input (for a complete overview see 5.2.1). If we do not take the institutional actors into account we found the following:

Least willing to share		**Most willing to share**	
Disclosure Dimension	Type of input	Disclosure Dimension	Type of input
Political attitude	Private Fb messages	Socio-demographic	Fb Events

Financial information	Pictures	Location	Likes

Due to the small number of respondents the differences between the pieces of information are rather insignificant so we do not present the full ranking here. If we compare these results with the results of the quantitative research track presented in 4.2.2 and Figure 13, we found some similarities as, the professional information (which also questioned income level in the quantitative track) is ranked second in the types of information our respondents preferred to keep private. We also found that location information and demographics ranked lower in the quantitative research track, and thus these respondents were also more willing to share this information. A side note should be made to this comparison since the qualitative research track always kept the institutional actors in mind.

6. Conclusion and next steps

This report provides insights in how users valuate their personal online information. The results were gathered through both a quantitative and qualitative research track. In the quantitative research track we had 182 participants voluntarily providing their Facebook data and filling in a survey with questions relating to the eight disclosure dimensions distinguished in WP 6. The questionnaire gave us insights into the perceived sensitivity of the disclosure dimensions and the necessity to keep the different disclosure dimensions private. It was found that health information, professional information and personality traits were perceived as the most sensitive pieces of information in general. Respondents also deemed it important to keep these types of information private. The participants were also asked if they thought the disclosure dimensions could be predicted from looking at the information at their Facebook profile. They rightly thought that demographics were most easy to predict. Political views are easier to predict than our respondents thought, the same is true for employment status and income level.

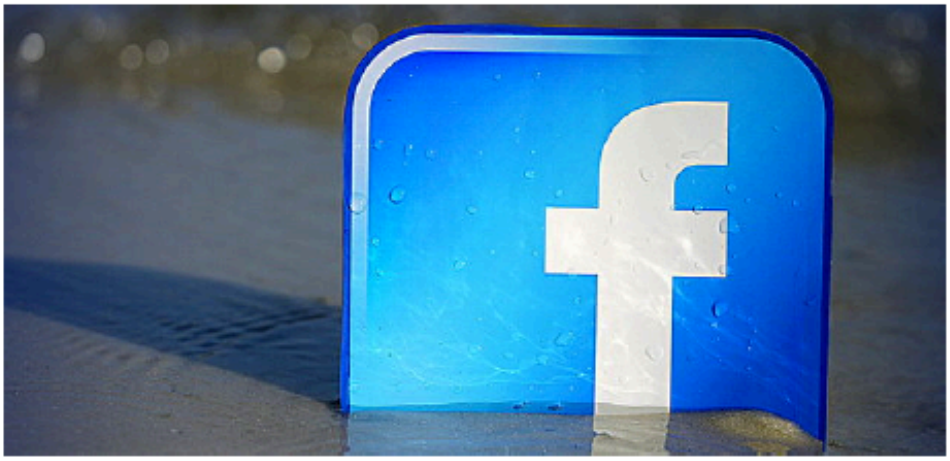
In the qualitative research track we then focused if the willingness to share personal information differed if the institutional actor reasoning on the data changed. We found in general that respondents were less willing to share their information with the insurance company and advertising agency, they were a little more willing to share the information with the government and an academic research organisation. Reasons for sharing information were that the information should be relevant for the workings of the organisation, not harmful to the data subject and it came from the right sources (e.g. LinkedIn data for (future) employers).

We also look at the attitudes of our respondents towards the different ways of data collection: volunteered, observed and inferred information. As expected people agreed more with the gathering of volunteered information, they became more critical towards the practice of observed information, since they feel there is nothing they can do to hide these kind of information as there is for example no proper alternative to browsing online for information. The attitude towards inferred information was in general very bad, because inferences could be wrong and you have no way in changing this since you do not know how you are being categorized.

As a next step in this process we need to analyse the Q sorts more in depth to see if we can distinguish different user groups with regard to willingness to disclosure personal information online. This can be backed up by a further analysis of the survey answers.

7. Annex

7.1. Invitation to the DataBait Research Tool



Heb jij controle over jouw informatie?

Hallo,

Bij **Facebook** verdienen ze een aardig centje aan het delen van **jouw informatie** met bedrijven en organisaties, die daar heel wat mee kunnen.

Onze collega's van het EU USEMP project vinden echter dat **iedereen recht heeft op zijn privacy**, ook online. Met hun **DATABAIT ONDERZOEK** willen deze wetenschappers daarom een tool ontwikkelen die de controle over wat er gebeurt met jouw persoonlijke data terug in jouw handen legt.

Maar ze kunnen dit niet alleen en vroegen ons daarom hun oproep voor deelnemers te verspreiden. Ze zoeken mensen

- 18 jaar of ouder,
- met een Facebook account,
- een basiskennis Engels,
- die bereid zijn de DataBaitapplicatie tijdelijk toegang te verlenen tot hun Facebookaccount*

**Alle verzamelde data wordt anoniem verwerkt en met de nodige zorgvuldigheid behandeld door de wetenschappers achter dit project.*

Meer informatie over hoe je kan helpen, vind je HIER.

Door deel te nemen, maak je kans op één van de 10 Fnac-bonnen ter waarde van €25.

7.2. Survey Quantitative Research Track

DataBait Research - Questionnaire

Dear participant,

Thank you for your participation in this research.

Filling in this questionnaire will take approximately 15 minutes. We'll ask you some questions about subjects such as your political attitude and brand preferences. Your responses will allow us to investigate if you reveal these types of information to third parties knowingly or unknowingly on your Facebook profile. In this way, we will be able to better warn you in the future when you share information that you prefer to keep private.

Of course you have our guarantee that we will protect your data with the highest concern.

If you need some more information, please don't hesitate to contact our researcher Tom: tom.seymoens@iminds.be

Set up of this questionnaire

On the following pages we will ask you short questions about the following subjects:

- Basic demographic information
- Professional and financial information
- Relationship information
- Religion
- Personality traits
- Sexual orientation
- Political attitude
- Health and condition
- Location information and holiday preferences
- Brand preferences and interests

Please be aware, that if you find a question too intimate you always have to possibility to skip it, by clicking 'no answer'.

1. **Basic Demographic Information**
- What is your Gender?
- What is your year of birth?
- What is your nationality?
- What is your country of residence?
- Do your parents or grandparents have origins from a different region?
- What is the highest degree or level of school you have completed?

2. **Feedback on Questions Demographic Information**
- How sensitive do you find the information you had to reveal?
- Do you think information on your Facebook profile reveals this information? Either because you yourself have put it online or it could be inferred from a combination of posts?
- How important is it for you that this type of information about you remains private?

3. **Professional and Financial Information**

- What is your employment status?
- Please indicate on the following scale where your monthly income level (before taxes) is situated. (approximately)

4. **Feedback on Questions Professional and Financial Information**

- How sensitive do you find the information you had to reveal?
- Do you think information on your Facebook profile reveals this information? Either because you yourself have put it online or it could be inferred from a combination of posts?
- How important is it for you that this type of information about you remains private?

5. **Relationship Information**

- What is your relationship status?
- What is your living situation?

6. **Feedback on Questions Relationship Information**

- How sensitive do you find the information you had to reveal?
- Do you think information on your Facebook profile reveals this information? Either because you yourself have put it online or it could be inferred from a combination of posts?
- How important is it for you that this type of information about you remains private?

7. **Religion**

- What is your religious stance?
- Do you actively practice this religion?

8. **Feedback on Questions about Religion**

- How sensitive do you find the information you had to reveal?
- Do you think information on your Facebook profile reveals this information? Either because you yourself have put it online or it could be inferred from a combination of posts?
- How important is it for you that this type of information about you remains private?

9. **Personality Traits**

- I see myself as someone who: (Denissen, Geenen, van Aken, Gosling, & Potter, 2008b)

No.	Dutch Translation	English Original
19	Zich veel zorgen maakt	Worries a lot
14	Gespannen kan zijn	Can be tense
9(r)	Ontspannen is, goed met stress kan omgaan	Is relaxed, handles stress well
39	Gemakkelijk zenuwachtig wordt	Gets nervous easily
24(r)	Emotioneel stabiel is, niet gemakkelijk overstuur raakt	Is emotionally stable, not easily upset
34(r)	Kalm blijft in gespannen situaties	Remains calm in tense situations
4	Somber is	Is depressed, blue
29	Humeurig kan zijn	Can be moody
1	Spraakzaam is	Is talkative
21(r)	Doorgaans stil is	Tends to be quiet
16	Veel enthousiasme opwekt	Generates a lot of enthusiasm
36	Hartelijk, een gezelschapsmens is	Is outgoing, sociable
6(r)	Terughoudend is	Is reserved
31(r)	Soms verlegen, geremd is	Is sometimes shy, inhibited
11	Vol energie is	Is full of energy
26	Voor zichzelf opkomt	Has an assertive personality
40	Graag nadenkt, met ideeën speelt	Likes to reflect, play with ideas
25	Vindingrijk is	Is inventive
30	Waarde hecht aan kunstzinnige ervaringen	Values artistic, aesthetic experiences
5	Origineel is, met nieuwe ideeën komt	Is original, comes up with new ideas
15	Scherpzinnig, een denker is	Is ingenious, a deep thinker
20	Een levendige fantasie heft	Has an active imagination
10	Benieuwd is naar veel verschillende dingen	Is curious about many different things
44	Het fijne weet van kunst, muziek, of literatuur	Is sophisticated in art, music, or literature
41(r)	Weinig interesse voor kunst heeft	Has few artistic interests
35(r)	Een voorkeur heeft voor werk dat routine is	Prefers work that is routine
3	Grondig te werk gaat	Does a thorough job
28	Volhoudt tot de taak af is	Perseveres until the task is finished
18(r)	Doorgaans geneigd is tot slordigheid	Tends to be disorganized
23(r)	Geneigd is lui te zijn	Tends to be lazy
13	Een werker is waar men van op aan kan	Is a reliable worker
33	Dingen efficiënt doet	Does things efficiently
38	Plannen maakt en deze doorzet	Makes plans and follows through with them
43(r)	Gemakkelijk afgeleid is	Is easily distracted
8(r)	Een beetje nonchalant kan zijn	Can be somewhat careless
32	Attent en aardig is voor bijna iedereen	Is considerate and kind to almost everyone
17	Vergevingsgezind is	Has a forgiving nature
7	Behulpzaam en onzelfzuchtig ten opzichte van anderen is	Is helpful and unselfish with others

Dutch Translation	English Original
Snel ruzie maakt	Starts quarrels with others
Soms grof tegen anderen is	Is sometimes rude to others
Koud en afstandelijk kan zijn	Can be cold and aloof
Mensen over het algemeen vertrouwt	Is generally trusting
Geneigd is kritiek te hebben op anderen	Tends to find fault with others
Graag samenwerkt met anderen	Likes to cooperate with others
% Variance (rotated)	

10. Feedback on Questions Personality Traits

- How sensitive do you find the information you had to reveal?
- Do you think information on your Facebook profile reveals this information? Either because you yourself have put it online or it could be inferred from a combination of posts?
- How important is it for you that this type of information about you remains private?

11. Sexual Orientation

- What is your sexual orientation?

12. Feedback on Questions Sexual Information

- How sensitive do you find the information you had to reveal?

- Do you think information on your Facebook profile reveals this information? Either because you yourself have put it online or it could be inferred from a combination of posts?
- How important is it for you that this type of information about you remains private?

13. Political Attitude

- How would you describe yourself concerning your ideology? (on a spectrum from left to right)

14. Feedback on Questions Political Attitude

- How sensitive do you find the information you had to reveal?
- Do you think information on your Facebook profile reveals this information? Either because you yourself have put it online or it could be inferred from a combination of posts?
- How important is it for you that this type of information about you remains private?

15. Health Factors and Condition

- In general how would you say your health is?
- Please read following statements concerning your smoking behaviour and tick the box next to the statement that describes you best.
- During the last 12 months, how often did you usually have any kind of drink containing alcohol?
- Please indicate next to each substance if you have used it in the last 6 months.
- What is your height? (in centimeters)
- How much do you weigh? (in kilograms)
- Do you suffer from a chronic disease? if so please specify.

16. Feedback on Questions Health Factors and Condtition

- How sensitive do you find the information you had to reveal?
- Do you think information on your Facebook profile reveals this information? Either because you yourself have put it online or it could be inferred from a combination of posts?
- How important is it for you that this type of information about you remains private?

17. Location Information

- In what city is your home located?
- In what city is your work located?
- Where did you spend your last holidays?

18. Feedback on Questions Location Information

- How sensitive do you find the information you had to reveal?
- Do you think information on your Facebook profile reveals this information? Either because you yourself have put it online or it could be inferred from a combination of posts?
- How important is it for you that this type of information about you remains private?

19. Consumer Information and Interests

- Please pick from the list the 5 favourite activities/interestsyou pursue in your spare time.
- What are your favourite brands that come to mind in the following categories:
    - Arts & Entertainment
    - Automotive Industry
    - Health & Fitness

- Food & Drinks
- Clothing & Fashion
- IT & Computing
- Travel
- Media & News
- Music
- Toys
- Other

20. Feedback on Questions Consumer Information and Interests

- How sensitive do you find the information you had to reveal?
- Do you think information on your Facebook profile reveals this information? Either because you yourself have put it online or it could be inferred from a combination of posts?
- How important is it for you that this type of information about you remains private?

7.3. Information Deletion Exercise Example (Advertising Agency)

HET RECLAMEBUREAU ZIET:

. Vrijwillig vrijgegeven informatie

- Nora werkt momenteel bij Pendant Publishing als copywriter
- Nora behaalde in 2004 haar masterdiploma in de journalistiek na deliberatie aan de universiteit van Gent
- Nora's moedertaal is Nederlands, ze spreekt ook vloeiend Frans en heeft een basiskennis van het Indisch
- Nora is getagd in een paar foto's waarop ze rookt. Nora is dus een roker
- Nora woont in de Pastoor Campensstraat 130, 2170 Merksem, België
- Nora's GSM nummer is 0471 33 03 31
- Nora Shankar en Carola Koster zijn in 2010 gaan samenwonen
- Nora melde onlangs in een status dat haar vrouw Carola Koster een nieuwe Toyota Auris heeft gekocht
- Nora deelt vaak artikels uit Het Laatste Nieuws over politiek zoals 'Overwinning NVA zal land doen beven'. Nora is geen aanhanger van de NVA
- Nora is fan van verschillende auteurs van gedichten op Facebook
- Nora is lid van de Facebook groep: 'Boeddhistische gemeenschap Antwerpen'
- Nora's Sleep Cycle applicatie toont aan dat Nora veel te onvast en weinig slaapt
- Nora's Runtastic applicatie toont aan dat Nora geregeld gaat lopen, maar ook dat ze een afwijkende hartslag heeft

. Geobserveerde informatie

- Nora heeft een Apple computer die in het Nederlands staat en inlogt vanuit Merksem
- In de cookies van Nora staat dat ze naar de volgende websites is geweest:
  - Scheidingsconsulenten.be
  - Twitter
  - De website van Paul Simon
  - Amazon.com en Zalando.be
  - Indiaweb.nl (actualiteit en cultuur van India)
  - Op google zat te zoeken naar een nieuw paar, rode sportschoenen
  - Immoweb.be
  - Rendez-vous.be
  - Het opzoeken van oorzaken voor slapeloosheid op dokteronline.com

Yoga-meditatie.org

. Afgeleide informatie

- De data vanuit Nora's loop en slaap applicatie en haar status als roker voorspellen dat ze 90% kans loopt om een hartfalen te krijgen binnen het jaar
- Uit de verschillende fan pagina's mbt gedichten en haar online zoekgedrag naar yoga en spiritualiteit kan worden afgeleid met 91% zekerheid dat Nora een praktiserend Boeddhist is
- De informatie van haar slaap applicatie en haar zoektocht naar oorzaken van insomnia kunnen er op wijzen dat Nora een onrustig persoon is (56%)
- Door analyse van de door haar bezochte websites (immoweb.be, scheidingsconsulenten.be en rendezvous.be) kan worden afgeleid dat Nora zal scheiden binnen de twee jaar (81%)
- Uit haar basis kennis Indisch, haar vreemd-klinkende achternaam (Shankar) en haar interesse in India kan worden afgeleid met 93% zekerheid dat Nora van Indiase afkomst is
- Uit haar job als copywriter, de aankoop van een nieuwe Toyota Auris en de buurt waarin ze woont kan worden afgeleid dat het gezin Shankar-Koster zich in de middenklasse bevindt.

8. Bibliography

Backstrom, L., & Kleinberg, J. (2014). Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook (pp. 831–841). ACM Press. http://doi.org/10.1145/2531602.2531642

Bucher, T. (2012). Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society*, *14*(7), 1164–1180. http://doi.org/10.1177/1461444812440159

Burke, M., Marlow, C., & Lento, T. (2010). Social network activity and social well-being. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1909–1912). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1753613

Creese, S., Goldsmith, M., Nurse, J. R., & Phillips, E. (2012). A data-reachability model for elucidating privacy and security risks related to the use of online social networks. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on* (pp. 1124–1131). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6296102

Denissen, J. J. A., Geenen, R., van Aken, M. A. G., Gosling, S. D., & Potter, J. (2008a). Development and Validation of a Dutch Translation of the Big Five Inventory (BFI). *Journal of Personality Assessment*, *90*(2), 152–157. http://doi.org/10.1080/00223890701845229

Denissen, J. J. A., Geenen, R., van Aken, M. A. G., Gosling, S. D., & Potter, J. (2008b). Development and Validation of a Dutch Translation of the Big Five Inventory (BFI). *Journal of Personality Assessment*, *90*(2), 152–157. http://doi.org/10.1080/00223890701845229

European Commission. (2015). *Special Eurobarometer 431 "Data Protection"* (Eurobarometer No. 431) (p. 220). European Union.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37.

Heyman, R. (2015, November). *Facebook & users: who is using who? A material semiotic approach to the irreversibilisation of Facebook as a case of lifeworld colonisation by social media*. Vrije Universiteit Brussel, Brussel.

Heyman, R., De Wolf, R., & Pierson, J. (2014). Evaluating social media privacy settings for personal and advertising purposes. *Info*, *16*(4), 18–32. http://doi.org/10.1108/info-01-2014-0004

Heyman, R., & Pierson, J. (forthcoming). Social media, delinguistification and colonisation of lifeworld: changing faces of Facebook. *Social Media + Society*, 12.

Heyman, R., & Van Dijk, N. (2013). *D3.3.1: Report on differences between user and legal perspective on privacy and profiling* (Deliverable, EMSOC Project) (p. 42). iMinds-SMIT, LSTS.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive scale emotional conta- gion through social networks. Proceedings of the National Academy of Sciences, 111, 8788–8790. doi:10.1073/ pnas.1320040111

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. E. Bijker & J. Law (Eds.), Shaping technology/building society: Studies in sociotechnical change (pp. 225–258). Cambridge Mass.: MIT Press.

A. Popescu, M. Hildebrandt, S. Papadopoulos, G. Petkos, Y. Kompatsiaris, L. Claeys, T. Seymoens, D. Lund, T. Michalareas, T. Kastrinogiannis, E. de Vries, N. van Dijk, J. Pierson, A.M. Padyab, E. Gadeski, H. Le Borgne, “User **Empowerment** for Enhanced Online Presence Management – Use Cases and Tools”, Amsterdam Privacy Conference 2015, pages 23-26, 8 October 2015, Amsterdam

Petkos, G., Papadopoulos, S., & Kompatsiaris, Y. (2015). Statistics for pre-pre-pilot data.

G. Petkos, S. Papadopoulos, Y. Kompatsiaris, “PScore: Enhancing Privacy Awareness in Online Social Networks“, International Workshop on Multimedia Forensics and Security (MFSec) held as part of the International Conference on Availability, Reliability and Security (ARES), 2015

Wang, Y.-C., Burke, M., & Kraut, R. E. (2013). Gender, topic, and audience response: an analysis of user-generated content on facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 31–34). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2470659