D2.3

Content and dataset specifications

v1.3 2016-03-01

Georgios Petkos (CERTH), Symeon Papadopoulos (CERTH), Adrian Popescu (CEA), Mireille Hildebrandt (iCIS), Noel Catterall (HWC), Yiannis Kompatsiaris (CERTH)

This deliverable looks at the variety of data that is used during the development of USEMP tools and that is foreseen to be handled by the USEMP system once it is operational. First, a number of data requirements are identified from the requirement analysis of the USEMP system. Apart from thoroughly presenting the actual data to be handled during system operation, a number of external datasets that are used to develop and evaluate the modules of the USEMP system are also presented. Moreover, the data storage and flow through the system are discussed. The deliverable also links to the legal analysis conducted in WP3.

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 2	Requirements and Use Case Analysis
Deliverable lead org.	CERTH
Deliverable type	Report
Authors	Georgios Petkos (CERTH) Symeon Papadopoulos (CERTH) Adrian Popescu (CEA) Mireille Hildebrandt (iCIS) Noel Catterall (HWC) Yiannis Kompatsiaris (CERTH)
Reviewers	Theodoros Michalareas (VELTI) Tom Seymoens (iMinds)
Version	1.3
Status	Final
Dissemination level	PU
Due date	2015-03-31
Delivery date	2015-04-19 (revised 2016-03-01)

Version	Changes
0.1	First outline ToC by CERTH
0.2	Material added in Chapters 2 and 3 by CEA
0.3	Material added in Chapters 1, 2 and 3 by CERTH
0.4	Additions in Chapters 1-5 by CERTH
0.5	Revisions and additions in all chapters by iCIS
0.6	Revisions in Chapters 1-4 by CERTH

0.7	Revisions in Chapters 5 and 6 by CERTH, submitted for internal review
0.8	Additions in Chapter 5 by HWC. Revisions by CERTH to cover comments from the first round of reviews.
0.9	Edits from HWC.
1.0	Pre-final version incorporating final refinements and amendments.
1.1	Final version incorporating comments from iCIS
1.2	Updated version addressing comments from the second annual review
1.3	Minor corrections after second round of proof-reading

Table of Contents

1.Introduction

The purpose of this deliverable is to describe the data that will be used during the development of USEMP tools and that will be handled by the USEMP system once it is operational. The description of the handled data has been compiled based on the system requirements. In particular, the focus has been on functional requirements, as presented in Section 4 of D2.2. This set of functional requirements includes a list of functions each of which is realized by a distinct software module of the USEMP platform. Each module takes as input some (raw) data and produces some derivative data as output, and is thus linked to a set of data requirements. In most cases, input data of these modules include various types of data and content collected from Online Social Networks (OSN), as well as web browsing behaviour data. Output data represent different types of information, some of which relate to various attributes inferred about a user's profile. Therefore, the USEMP data mostly consist of data that come from a) monitoring OSN presence and web behaviour, and b) from producing various results about the profile of a user. In addition, there is a number of external datasets, not directly handled by the USEMP system, but which are being used in order to train, tune or evaluate different modules, which are also presented in this report. Importantly, the report also discusses a number of data management issues, such as storage, as well as the legal obligations that apply to storing and otherwise processing the data.

This document is related to other deliverables and further elaborates on their respective outcomes. For instance, various aspects of the handled data are presented in deliverables produced from WPs 5 and 6: D5.1, D5.2, D5.3, D6.1 and D6.2. More specifically, information about external datasets that are used to train inference mechanisms has been presented in these deliverables, albeit the focus on these deliverables has not been on the datasets, it was rather on the inference mechanisms. Moreover, an early discussion of the inputs and outputs of the modules has been presented in D7.1 Architecture Design. Additionally, some information related to the results produced by the USEMP system was presented in D6.1, in which the USEMP scoring framework has been discussed. Also, as already mentioned, we make a link to the requirements that are presented in D2.2 and in addition to the set of social requirements presented in D4.1. Finally, this document is related to the legal research in WP3 and in particular to the interface between legal and technical research that is presented in D3.4 and will be further investigated in D3.6-D3.9.

This document is structured as follows. In Chapter 2 we look at the requirements of the USEMP platform and identify a number of data requirements. We observe that these data requirements can be organized in two classes. The first class is related to a number of modules specified by the functional requirements, whereas the second class is about the overall USEMP user profile. Then, in Chapter 3 we examine in more detail at the external datasets used to train and fine-tune the USEMP modules, and then we look at some of the data that have been produced internally during the pre-pilots and the system operation and which will be used for further development of the modules. Subsequently, in Chapter 4 we move our focus to the USEMP user profile. As will be described, the USEMP user profile consists of data that directly represent the presence of the user on the OSN and their browsing behaviour, and of data produced by the USEMP modules. Then, Chapter 5 examines various data management issues, such as the description of the flow of data though the system, as well as storage issues. Finally, Chapter 6 concludes the deliverable.

1.1. Terminology

Before proceeding to the core content of the deliverable, we clarify some terminology issues that are important for the discussion that will follow.

The terms *privacy*, *personal data* and *sensitive data* are part of our common vocabulary, used in a variety of ways by different people depending on context and personal inclination. Notably data scientists, digital security experts, lawyers and social scientists have different understandings of these terms. We note that insofar as such terms are legal terms, they have legal effect, which means that once a practice, operation or activity is qualified as such they generate legal rights and obligations. Within the USEMP project, research is conducted to detect privacy perceptions of end-users of OSNs and to figure out which of their volunteered, observed and inferred data they qualify as sensitive. Legally speaking, the term 'sensitive data' or 'personal data', however, has a more precise meaning and once data is qualified as such this has legal effect. The legal effect means that a bundle of legal rights applies to the end-user and a bundle of legal obligations applies to the service providers (i.e. the USEMP Consortium Partners). We need to prevent confusion over whether terms like privacy, personal data or sensitive data are intended as an indication of how end-users perceive specific data (usage) or as referring to the legal qualification of an activity as 'personal data processing', or even as 'processing of sensitive personal data'. For this reason we have developed the following strategy:

- Privacy as perceived by end-users is framed as either perceived privacy or as non-disclosure (if that is what is actually at stake);
- Whenever issues of data protection law are at stake we frame them in terms of privacy in the legal sense, as data in the legal sense or as sensitive data in the legal sense. If confusion is out of the question we simply speak of personal data or sensitive data.

Clearly there is overlap between the legal qualification of privacy and data protection on the one hand and the perception of privacy lost and gained on the other. The point is, however, to acknowledge that rights and obligations have been attributed by the democratic legislator and often by constitutional legislators, to protect people against the need to trade their freedoms. In that sense these rights and obligations do NOT depend on individual preferences. For instance, the obligations of data minimisation and purpose limitation cannot be nullified on the basis of consent; even if consent is given, only those data may be processed that is necessary to achieve the specified purpose of processing, and only as long as the purpose is not exhausted. The objective of USEMP is to provide a more level playing ground for end-users to exercise their rights, by informing the end-users of potential inferences made on the basis of some of their data points or machine-readable behaviours. Though we will analyse their responses to the profile transparency that is provided by the DataBait tools, the need for a level playing field is not a matter of individual preferences. It is – on the contrary – a precondition to develop and act upon such preferences.

2.USEMP Data Requirements

2.1. Overview

In this chapter we look into the data requirements of the USEMP system. We identify two types of data requirements and we organize the discussion around them. The first type of data requirements is related to the set of functional requirements, as listed in Section 4 of D2.2. These functional requirements dictate the development of a number of services, each of which is actually implemented by a specific module. The development of each of these modules implies a set of data requirements. Most of the modules take as input specific types of data related to the OSN presence or browsing behavior of a USEMP user and produces some information related to the user's profile. It should also be noted that most of these modules, which have been developed within WPs 5 and 6, have been trained or evaluated using a number of external datasets and will be further trained and evaluated with data from the pre-pilots and the actual operation of the system. The second type of data requirements is related to the fact that an overall USEMP user profile needs to be maintained by the system. This user profile contains unprocessed data resulting from directly monitoring the user's behavior, in the OSNs or their web browsing behavior (observed data) and data produced by the data driven modules (inferred data).

This overall scheme is displayed in Figure 1: starting from the overall set of requirements, we identify the two types of data requirements and then we identify specific sets of data that are used by the system. In addition, there exist a number of storage issues that apply to the specific data used by the system as well as a number of data usage issues that are related to the overall USEMP platform. Finally, the processing of all data must be compatible with the legal framework of data protection and – in some cases – with Intellectual Property rights. The legal analyses with regard to DataBait tools have been performed in D3.1-D3.4; those for the external datasets will be performed in the D3.6-D3.9.

It should be noted that this organization of the USEMP data is not mutually exclusive, i.e. there is some overlap between some parts of the presented data. More specifically, the development data produced during system operation partly overlaps with the data collected from the OSN. Nevertheless, this organization of the data has been selected because it focuses on the two major constituents of the USEMP platform: the data-driven modules and the USEMP user profile.

In this chapter we will introduce the two types of data requirements and in the next two we will look in more detail at the specific data that cover these requirements. Then, in Chapter 5 we will look at data usage, and storage issues. Throughout the deliverable we will refer to potential legal implications that will be further developed in the context of WP3.

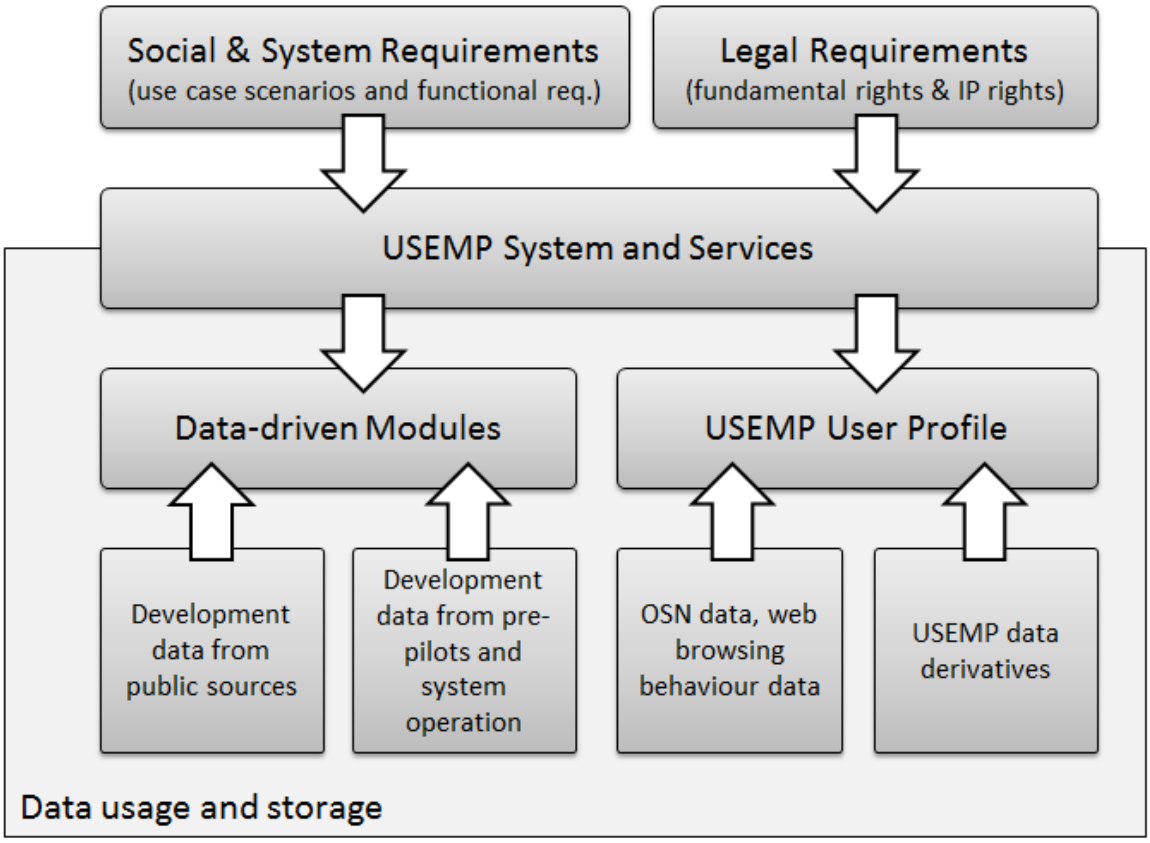

Figure 1. Overview of the USEMP data requirements and handled data.

2.2. Data-driven modules

We start the definition of data requirements by examining the set of functional requirements (Section 4 of D2.2). Each functional requirement is associated to a data-driven module. The set of USEMP modules is listed in Table 1. For each of the modules, the set of inputs and outputs is also listed and this effectively provides the first set of data requirements. Please note that pointers to relevant deliverables are also provided.

In the next section, we look at the second set of data requirements, i.e. those related to the USEMP user profile. In the next chapter we will come back to the data driven modules and will examine in more detail specific instances of datasets that can be used to – where applicable - tune or evaluate the data driven modules.

We summarize the above in Table 1.

Module	**Related deliverable**	**Input**	**Output**
Face recognition	D5.2	Images	Detected faces (number and location)
Logo recognition	D5.2	Images	Detected logos (number and location)
Multimodal concept detection	D5.3	Images	Concepts related to (perceived) privacy

Text similarity	D5.1	Textual posts	Concepts related to (perceived) privacy
Concept detection	D5.1	Textual posts	Concepts and entities related to (perceived) privacy
Opinion mining	D5.4	Textual posts	Polarity of text (positive / negative / neutral)
Location detection	D5.1 / D5.2	Images / posts / multimodal items	Set of locations that are present in the user's data
Large scale visual concept recognition	D5.2	Images	Concepts related to (perceived) privacy
Personal attribute behavioral detection	D6.1	Likes	User attributes related to (perceived) privacy
Topic based attribute detection	D6.1	Textual posts	User attributes related to (perceived) privacy
Network based attribute detection	D6.1	Friendship network and privacy related attributes for some of the friends	User attributes related to (perceived) privacy
Disclosure scoring framework	D6.1	Results of inferences, OSN presence data	Disclosure scores
Personal data value scoring framework	D6.1	Results of inferences, OSN presence data	Personal Data value scores
Disclosure settings assistance framework	D6.2	Disclosure scores, settings, all OSN presence data, etc.	Disclosure settings suggestions and settings

Table 1. Summary of data driven modules, each of them is related to a number of data requirements which is defined in terms of the input and output of the module.

2.3. USEMP User Profile

Having presented the set of functional requirements, the corresponding modules and the related data requirements, we now turn to the USEMP user profile. As mentioned, the USEMP user profile consists of two types of data. The first is the raw data that is either retrieved from the OSN (volunteered data) or is collected by monitoring the web browsing behavior of the user (observed data), while the second is data produced by the USEMP system after analyzing the raw OSN data and the browsing behavior data (inferred data).

OSN presence data contain data such as the status updates or the multimedia items posted by the user and any explicit profile information that is provided by the user. Web browsing behavior data consist of the set of visited URLs. Produced information contains the outputs of most of the inference modules, such as detected privacy attributes and values.

The data produced by the system contains information related to a number of privacy related attributes, such as the user's age, location, etc. All this information is organized in a structure

that is related to the disclosure scoring framework, which was presented in D6.1. In short, the scoring framework organizes the perceived privacy related data of a user in a hierarchical and semantic manner: a number of perceived privacy dimensions is identified (e.g., demographics, religious views, etc.), each of which is related to a number of attributes (such as those mentioned before, e.g. age, location, etc.) and each attribute can take a number of values. Values are linked to specific OSN data that support it either directly or through inference mechanisms. At the same time, a number of disclosure scores are produced for each part of the hierarchy. The perceived-privacy scoring framework is the main tool that is used by the USEMP platform to enhance the awareness of users in relation to what they consider privacy-related information, but at the same time it is a major component of the internal representation of the profile of a USEMP user.

Finally, it is useful to link the USEMP user profile data to the OSN data taxonomy that was introduced in (Schneier, 2010) and is also discussed in D6.1. This identifies six categories of OSN data:

- *Service data.* This is the set of data that a user explicitly provides to the OSN service.
- *Disclosed data.* This includes the content (messages, status updates, photos, etc.) posted by the user to their own page.
- *Entrusted data.* This is the content posted by the user to the page of another user.
- *Incidental data.* This is the content posted about the user by some other user.
- *Behavioural data*. This type of data includes the actions of the user in the OSN.
- *Derived data*. This is data about a user that may be derived from other types of data.

The first five categories of data, is what we actually refer to as “OSN presence data”, whereas the sixth category contains what we term “USEMP data derivatives”.

3.Data related to USEMP Data-driven Modules

Having sketched the basic data requirements of the USEMP modules, we now proceed to examine in more detail the specific data that is used to develop and tune the USEMP data-driven modules. As already mentioned, there are two types of relevant datasets: a) datasets that have been obtained from external sources, and b) data that is produced by the operation of the USEMP system and the pilots.

3.1. External datasets

3.1.1. MyPersonality

Dataset description

The MyPersonality dataset[1] resulted from a Facebook application with the same name that allowed users to take psychometric tests. The resulting data has been made available to the scientific community, in an anonymized form, in order to advance OSN-related research[2]. Importantly, the dataset contains data related to various privacy-related attributes for a significant part of the users. This includes information about the demographics details of users (available for 4,282,857 users), their political beliefs (available for 330,892 users) and their religious beliefs (available for 330,781 users). Demographics details that are available include the gender, birthday, age, and relationship status of users. However, it should be noted that not all types of demographics information is available for all 4,282,857 users for which some type of demographics information is available. For instance, the gender of most users (4,202,360) is available; in contrast, information on sexual orientation is available for far fewer users (1,168,456). The dataset also includes the anonymized likes of 253,705 users. It is also important to note that the likes of users that belong to some class (e.g. male / female) are available only for a subset of the users that we know that belong to that class.

Dataset usage in USEMP

We have used this dataset for experimenting with methods that are able to predict personal attributes of users based on their OSN behavioural data (likes in particular) and topics extracted from their OSN data. These experiments have been presented in D6.1. It should be noted though that, unfortunately, the MyPesonality dataset cannot be used for training modules that will be integrated to the USEMP system. There two reasons for this. The first is that the MyPersonality dataset is fully anonymised, i.e. both users and liked pages are replaced by numeric identifiers and therefore we cannot match the likes of USEMP users to those of the MyPersonality users. The second issue is that the dataset was constructed by surveying mostly English-speaking subjects, so it will not be suitable for use with the Belgian and Swedish users of USEMP. Thus, other datasets and/or the data collected during system operation and the early pilots are needed. Indeed, in D6.4 we further experimented with

[1] http://mypersonality.org/

[2] In the context of D3.6 and D3.9 we will investigate the anonymisation techniques employed and evaluate whether the data qualifies as anonymised data in terms of the EU legal framework (which obviously does not apply to US researchers). On the issue of de-anonymisation in large datasets see Paul Ohm, “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization”, UCLA Law Review 57 (2010): 1701–77. On the implications of the EU legal framework we will employ the Art. 29 WP Opinion 5/2014 on anonymisation techniques (WP216).

relevant prediction methods using data from the early pilots for training and testing. Nevertheless, data from the MyPersonality dataset has been very useful for drawing a number of conclusions about the predictability of various user attributes and the appropriateness of different prediction settings (e.g. type of classifier, feature selection techniques, etc.). More particularly, the following user attributes were examined when using only behavioural features. Please note that the number of users related to each attribute value is shown in parentheses and that the best classification accuracy achieved for each classification task (measured in terms of area under the receiver-operating characteristic curve) is also reported:

- Sexual preference: gay (21,592) / straight (441,935); accuracy: 86.39
- Marital status: single (992, 259) / married (291,944); accuracy: 70.92
- Political views: liberal (42,679) / conservative (35,706); accuracy: 83.01
- Religious beliefs: christian (170,802) / muslim (12,999); accuracy: 87.27

For more details on these experiments and the obtained results please see D6.1.

Legal considerations

We note that part of the data would be sensitive data in the legal sense, with a stricter regime of data protection law, if the data would enable identification. It is not obvious that an exception for scientific research would apply, notably not if the purpose of processing would be the commercial exploitation of the findings. Regarding licensing, an explicit permit needs to be obtained from the creators of the dataset and is granted based on its intended usage.

3.1.2. PicAlert dataset

Dataset description

The PicAlert dataset[3] was used in the study presented in (Zerr et al., 2012). In short, this study involves a series of experiments in which publicly available images uploaded to Flickr are classified as either “private” or “public”. The images have been annotated manually by users (not the original owners) and for each of them a number of textual features are available: title, description and tags. The authors also use a number of visual features extracted from the images: detected faces, color histograms, edge-direction coherence vectors, SIFT features, brightness and sharpness. They performed a number of classification tasks, each time using a different set of features. The dataset contained 31,010 images, 4,665 of which were labeled as private and 26,375 were labeled as public.

Dataset usage in USEMP

The PicAlert dataset is used for the development of the privacy settings assistance module. More specifically, it is used to train a classifier similar to the one presented in (Zerr et al., 2012); if the classifier predicts that an image is “private”, then the user will be warned about possible disclosure risks stemming from sharing the image. Satisfactory results have been obtained in early experiments with the PicAlert dataset. In particular, a Break-Even Point (BEP) of 0.7529 has been achieved, competitive to that reported by (Zerr et. al., 2012) that was 0.78. For more details please see D6.2. It is worth noting though that the relevant work from D6.2 has since been extended and these more recent developments are discussed in D5.5, where a new dataset was collected specifically for the needs of the project and was

[3] http://l3s.de/picalert/#ustudydata

used to carry out the same task. This dataset has a similar form to the PicAlert dataset with the main difference being that we also know the characterization of images from specific users. The main outcome is that the perception of what types of images are characterized as "private" or "public" tends to vary between users and therefore personalization is required to better match individual users' needs. For more details please see D5.5.

Legal considerations

Regarding legal considerations, it is important to note that in the context of D3.7 and D3.9, the applicability of copyright protection on the relevant images will be investigated; insofar as the images concern identifiable natural persons the applicability of data protection law will be studied in D3.6. The PicAlert dataset is not licensed.

3.1.3. Location estimation dataset

Dataset description

This is the dataset used for the 2014 MediaEval Placing Task[4]. It consists of 5 million geotagged photos and 25,000 geotagged videos that are used for training, and 500,000 photos and 10,000 videos that are used for testing. The training and the test set are mutually exclusive with respect to the users who contributed the content (i.e., the users in the training set will be different from the users in the test set). Importantly, all photos and videos used in the benchmark have been taken from the YFCC100M dataset, hence they are available under the Creative Commons license.

Dataset usage in USEMP

This dataset is used to train and benchmark the accuracy of the location detection modules developed within D5.1 and D5.2 (based on textual and visual features respectively). The relevant module has already been integrated to the platform since the early pilots and is currently being improved. The accuracy of the method that was developed based on this dataset is satisfactory: indicatively, using only textual features the precision at a distance of 10 kilometres was 0.613 and using only visual features it was 0.032. It is also worth noting that the methods developed using this dataset have been used to successfully participate to the MediaEval Placing Task. The use of the dataset is discussed in D5.1, D5.2 and D5.4.

Legal considerations

Insofar as location data is processed that can be linked with an identifiable natural person the legal implications will be investigated in D3.6. The location estimation dataset comes with a Creative Commons license.

3.1.4. Kaggle community detection dataset

Dataset description

This dataset was originally used for the Kaggle challenge "Learning Social Circles in Networks"[5], a competition with the goal to split the friends of a number of (anonymized) OSN users into appropriate social circles. For each of those users, the dataset provides a list of their friends, anonymized Facebook profiles of each of those friends, and a network of

[4] http://www.multimediaeval.org/mediaeval2014/placing2014/
[5] http://www.kaggle.com/c/learning-social-circles

connections between them (i.e. their “ego network”)[6]. In total there are 110 ego networks and for 60 of those we have the hand-labelled communities provided by each user. Examples of profile features provided include the following: birthday, classes attended, degree attended, school attended, year school was completed, family name, gender, location, political views, religious views, work position, employer, etc.

Dataset usage in USEMP

This dataset is used for training purposes in the disclosure settings assistance module. More specifically, it is used to validate different graph clustering algorithms, as well as for training a model that assists in ego-network clustering, with the goal of grouping the friends of a user in meaningful “social circles”. These social circles are eventually to be used as audience sets in a disclosure policy defined by the user. Essentially, the automatic generation of social circles will allow the user to easily define sets of users to which disclosure of specific types of content should be allowed or not. A number of graph clustering algorithms have been tested with the dataset and the best score achieved (measured in terms of Normalized Mutual Information) was 0.633. More details about the relevant experiments can be found in D6.2.

It should be noted though that the actual integration of this functionality in the platform is restricted due to the fact that access to the full ego-network of a user is limited. More specifically, due to Facebook API limitations, it will be possible to only construct the part of the ego-network of a user that contains other friends of the user that also use DataBait.

Legal considerations

To the extent that the data is not anonymized this would concern a plethora of sensitive data in the legal sense, for which a strict regime applies. We will investigate this further in D3.6 and D3.9. The Kaggle dataset is not licensed.

3.1.5. Relevance- and diversity-based reranking dataset

Dataset description

This dataset was used in the 2014 Retrieving Diverse Social Images (RDSI) task of MediaEval (Ionescu et al., 2014). The task addressed the problem of result diversification in social photo retrieval. Participants (recruited by the organizers of the task - not associated with USEMP) were provided with an ordered list of up to 300 images returned by Flickr in response to a textual query for a specific Point of Interest (POI) and were asked to refine this list by providing a ranked list of up to 50 images that are both relevant and diverse representations of the query. Explicit definitions were provided for both relevance (e.g., artistically deformed photos are relevant, while photos that present an aspect of a POI that is not socially recognizable are not considered relevant) and diversity (e.g., different times of the day/year). The refinement and diversification process could be based on the information provided for each POI (Wikipedia page, up to five representative photos from Wikipedia, GPS coordinates), the metadata of the retrieved images (e.g., title, description, tags, GPS coordinates, etc.) as well as their visual content. During the task, participants were provided with an annotated development set of 30 queries (ground truth) - in order to build their approaches - as well as a test set of 123 queries - upon which they were evaluated. Ground truth consisted of relevance and diversity annotations provided by experts for all images of

[6] See note 2, the anonymization techniques used in the dataset of the Learning Social Circles in Networks will be tested in D3.6 and D3.9.

the test set. Specifically, each image was first labelled as either relevant or irrelevant and then visually similar relevant images were grouped together into clusters.

Dataset usage in USEMP

This dataset is used for benchmarking the method used for the relevance and reranking module that is used as part of the large scale visual concept recognition module. In particular, the relevance and reranking module is used to present to the user a ranked and diversified list of the images that are related to some concept. This ranked and diversified way of presentation will allow the user to effectively provide relevance feedback to the system, so that visual concept recognition results are improved. The method that was developed with the use of this dataset was also used to successfully to the RDSI task (achieved score, F1@20, was 0.631). Further details about the use of this dataset can be found in D5.3.

Legal considerations

Again, the applicability of data protection and copyright law on this set will be analysed in D3.6 and translated into evantual requirements in D3.9. The relevance- and diversity-based dataset comes with a creative commons license.

3.1.6. Wikipedia

Dataset description

The well-known online encyclopedia is developed collaboratively by volunteers and includes descriptions of a wide number of concepts in many languages. For instance, as of March 2015, the versions of Wikipedia in the most probable languages used by USEMP users include: 4.74 million concepts (articles) for English, 1.8 million for Dutch, 1.9 million for Swedish and 1.6 million for French. Equally important, the encyclopedia includes a large number of inter-lingual links and it is thus possible to correlate knowledge across languages and, whenever necessary, process multilingual content. Due to its richness and immediate availability, Wikipedia is one of the most useful resources for a wide spectrum of NLP tasks.

Dataset usage in USEMP

Wikipedia is used for developing the text similarity module. More specifically, building on initial work by (Bouamor et al., 2013), Wikipedia is exploited in order to develop domain representations that will help classify users' shared texts into privacy-related domains such as politics, health or religion. These classifications can be used alone to provide feedback about what a third-party can infer from raw texts or be integrated in the privacy scoring framework developed as part of D6.1. Essentially, we use Wikipedia in order to obtain representations of specific privacy concepts and use them to identify these concepts in the textual content that is posted by the users. It is important to note that we have downloaded English, Dutch, Swedish and French Wikipedia dumps in order to cover the most common languages that we expect to come across during system operation In the set of experiments that we carried out, the method that we developed manages to perform quite well (precision @10 was 0.468) as compared to a more classical approach (precision @10 was 0.212). For more details about the method and the experimental results please see D5.1.

Legal considerations

The usage of Wikipedia is controlled by special licensing terms, specific to Wikipedia that can be found in https://en.wikipedia.org/wiki/Wikipedia:Copyrights.

3.1.7. SentiWordNet

Dataset description

SentiWordNet (Baccianella et al., 2010) is a publicly available[7] lexical resource built to support opinion mining and sentiment analysis tasks. It is built on top of WordNet and includes a positivity/neutrality/negativity index to every WordNet synset. This index is automatically obtained through semi-supervised learning and a random walk for score refinement. For instance, the synsets with top positive scores include: *good, better_of_all, divine* or *superb*. Inversely, synsets with most negative connotation include: *abject, deplorable, bad or scrimy*. SentiWordNet is extensively used in the opinion mining frameworks and challenges, including the SemEval Sentiment Analysis in Twitter track[8]. The resource was developed for English and its adaptation to other languages, needed in USEMP, poses two important challenges: (1) the mapping of English WordNet toward other languages is incomplete and only a part of synsets are annotated and (2) although aligned in WordNet translations, the meanings of words varies across languages and the index associated to English words might be only partially accurate in other languages.

Dataset usage in USEMP

Although its usage is challenging, SentiWordNet remains a very useful resource and will be included in an adaptation of CEA's sentiment analysis tool (Marchand et al., 2013) that will be performed during the second iteration of T5.1 work. More particularly, the sentiment analysis module is going to work complementary to the text similarity module. That is, whereas the text similarity module will detect the presence of privacy related concepts to the textual content posted by the user, the sentiment analysis module will detect whether the user has a positive, negative or neutral attitude towards the concept.

Legal considerations

The applicability of copyright law on this set will be analysed in D3.6 and translated into eventual requirements in D3.9. The dataset comes with a Creative Commons license.

3.1.8. ImageNet

Dataset description

ImageNet (Deng et al., 2009) is a large-scale visual resource that is built by populating a significant part of the WordNet noun hierarchy with images. As of March 2015, the dataset contains over 14 million images depicting nearly 22,000 synsets (concepts)[9]. The relevance of the images was checked by human annotators and they generally provide an accurate illustration of concepts. ImageNet is publicly available and it has stimulated extensive research in large-scale image mining. For instance, a subset of 1,000 concepts is used in the popular ImageNet challenge (Russakovsky et al., 2014), of which the main objective is to evaluate object recognition and localization. Moreover, ImageNet is notably exploited to train state-of-the-art deep learning models (Sermanet et al., 2013; Jia, 2013).

[7] http://sentiwordnet.isti.cnr.it/
[8] http://alt.qcri.org/semeval2015/task10/
[9] The term *concept* is an established term in the multimedia analysis and computer vision research communities and is typically associated with a topic, entity, object or theme depicted in an image.

Dataset usage in USEMP

In USEMP, ImageNet content is mainly used in D5.2 to create a large number of image classifiers that are exploited to inform users about privacy-related insights that can be gained from an analysis of their shared images. Of particular interest is the reuse of an ImageNet subset of concepts that are privacy-related in order to build a privacy-oriented visual dataset. This usage can be either direct, through the presentation of statistics about the concepts detected in the users' images or integrated to the disclosure scoring framework developed as part of D6.1. While it includes a large number of concepts, ImageNet fails to cover popular items and notably: commercial artefacts (product and brand names), events (sports and cultural events) and persons (famous people). To overcome this limitation, USEMP will exploit a semi-automatic extension of ImageNet that illustrates Wikipedia concepts with Web images. This resource is currently developed as part of the FP7 MUCKE project and will be released under an open access license. It will be notably exploited to improve product and face recognition during the second iteration of T5.2.

Legal considerations

The applicability of data protection and copyright law on this set will also be analysed in D3.6 and translated into eventual requirements in D3.9. The licensing of the ImageNet dataset specifies that it can be used for non-commercial research.

3.1.9. Logo recognition datasets

Dataset description

FlickrLogos-32[10] is a publicly available dataset that includes manually checked images for 32 logos. The full set includes 8,240 images. There are three partitions of FlickrLogos-32: P1 (training set) includes 10 images per class and is used for training; P2 (validation set) includes 10 logo images per class and 3,000 distractor images; P3 (test set) includes 30 logo images per class and 3,000 distractor images.

Dataset usage in USEMP

The FlickrLogos-32 dataset is used for the development of the logo recognition module. Identification of logos in images shared by users effectively results in the identification of their consumer behavior. In the first iteration of T5.2, the dataset was used to test local image descriptor based implementations (Romberg et al., 2011). To take advantage of recent developments from the deep learning field, whose accuracy increases when large sets of data are available, the logo recognition dataset will be extended in two directions during the second iteration of T5.2. First, logos of popular products and brands will be added in to reach 500 items and be able to recognize a larger portion of the product or brand images shared on OSNs. Second, a much larger number of images will be collected from the Web. A part of the collected images will be annotated manually and then exploited in an image reranking setting in order to reduce the quantity of noisy images in the initial dataset collected from the Web. Our approach currently achieves an MAP score of 0.48; for more details please see D5.2.

Legal considerations

Use of the dataset is governed by the Flickr terms of use.

[10] Available at http://www.multimedia-computing.de/flickrlogos/ (accessed on 12/03/2015)

3.1.10. Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M)

Dataset description

The YFCC100M dataset (Thomee et al., 2015) is one of the largest publicly available multimedia collections. It was already exploited in challenges such as the MediaEval Placing Task 2014 and ACM Multimedia Grand Challenge on Event Detection and Summarization and is likely to become a standard collection in multimedia mining. It includes around 100 million Flickr images and videos, with associated metadata (such as: identifier, owner name, camera, title, tags, geographic coordinates), that were shared between 2004 and 2014. Due to its size and collective production, the dataset covers a large number of concepts and offers a collective snapshot of photographic practices.

Dataset usage in USEMP

While the YFCC100M dataset does not provide manually gathered relevance judgements for its items, it can still be useful in a number of multimedia related applications that are relevant for USEMP. For instance, it includes over 40 million geotagged images that cover locations all over the world and can be used in WP5 as a background collection for visual and multimedia content geolocation. YFCC100M can also be used to illustrate a wide variety of concepts and can thus constitute a valuable, though noisy, training set for visual concept modelling. In particular, given that all included images are shared under Creative Commons license, a subset of YFCC100M images that correspond to privacy-related concepts will be exploited in conjunction with an ImageNet subset to build a dataset focused on such concepts, created as part of T5.2 work.

Legal considerations

The applicability of data protection and copyright law on this set will also be analysed in D3.6 and translated into eventual requirements in D3.9. The dataset comes with a Creative Commons license.

3.1.11. SNOW dataset

Dataset description

The SNOW dataset (Papadopoulos et al., 2014) was put together for a topic detection competition. It consisted of a large number of tweets, around 1,100,000 for the development dataset and around 1,040,000 for the test dataset. The dataset also contained a set of ground-truth topics that were represented as sets of keywords.

Dataset usage in USEMP

The SNOW dataset has been used in order to train a module that predicts various privacy attributes for a user based on the concept of homophily. This is described in more detail in D6.1, but in short it uses the network of interactions (mentions) around a user and known privacy attributes of the user's network in order to predict the value of the privacy attribute for the user. The developed method (for more details please see D6.1) was tested on a network that was extracted from the SNOW dataset. The extracted network was built by taking into account mentions and replies. User labels corresponding to specific privacy dimensions (political opinion, religious beliefs and location) were generated for 13,000 users of this network (these were selected by computing PageRank on the graph and selecting the top ranking users). In order to get the labels for the selected users, we used the Twitter API to collect up to 500 Twitter lists that these users belong to. The list names and descriptions

were then tokenized, stop words were removed, and the remaining tokens were lemmatized. With manual inspection, some lemmas were removed and others were merged into a final list of labels. Using TF-IDF scoring, we computed a user-label matrix and then selected as "correct" labels, those that were associated with a user with a score equal to, or higher than the 75th percentile of normalized frequencies.

Legal considerations

The applicability of data protection and copyright law on this set will be analysed in D3.6 and translated into eventual requirements in D3.9. The SNOW dataset is not licensed.

3.1.12. Summary of external datasets

We summarize the above by listing in Table 2 the set of modules that will be developed and the corresponding external datasets that will be used for the development of each of them.

Module	Dataset	License
Face recognition	YFCC100M (Sec. 3.1.10)	CC license
Logo recognition	Logo recognition dataset (Sec. 3.1.9)	Flickr terms of use
Multimodal concept detection	ImageNet (Sec. 3.1.8)	Non-commercial research
Text similarity	Wikipedia (Sec. 3.1.6)	Wikipedia licensing terms
Concept detection	Wikipedia (Sec. 3.1.6)	Wikipedia licensing terms
Opinion mining	SentiWordNet (Sec. 3.1.7)	CC license
Location detection	Location estimation dataset (Sec. 3.1.3) YFCC100M (Sec. 3.1.10)	CC license CC license
Large scale visual concept recognition	Relevance- and diversity-based reranking dataset (Sec. 3.1.5) ImageNet (Sec. 3.1.8) YFCC100M (Sec. 3.1.10)	CC license Non-commercial research CC license
Personal attribute behavioral detection	My Personality (Sec. 3.1.1)	Usage permit required
Topic-based attribute detection	My Personality (Sec. 3.1.1)	Usage permit required
Network-based attribute detection	SNOW dataset (Sec. 3.1.11)	Non licensed
Disclosure scoring framework	Relevance- and diversity-based reranking dataset (Sec. 3.1.5)	CC license
Disclosure settings assistance framework	PicAlert (Sec. 3.1.2) Kaggle social circle dataset (Sec. 3.1.4)	Non licensed Non licensed

Table 2. Summary of external datasets used for each of the developed modules.

3.2. Pre-pilot data

Dataset description

Pre-pilot data consist of both the raw OSN presence data and browsing behavior data of the users that participated as well as questionnaire data that contain privacy related information about them. The questionnaire effectively provides the ground truth for the privacy related attributes that we consider. Given that the questionnaire has already been presented in D4.2, in the following, we examine in more detail the OSN presence data and then we examine the browsing behavior data that is collected. As discussed in D3.1, these data must be processed under the heading of the Data Licensing Agreement and any processing operation must be in conformity with applicable data protection law. The requirements for the processing of this data are summed up in D3.4. Further requirements follow from D3.2 and D3.3, as also stipulated in D3.4.

For the pre-pilots, OSN data is requested only from Facebook. Data about a user that is requested from Facebook by DataBait is listed in Table 3.

Category / permission	**Description**
public_profile	The public profile contains the basic information about a user. It includes the following sub-fields: • id • name • first_name • last_name • link • gender • locale • timezone • updated_time • verified
user_friends	This is the list of friends of the user that also use the DataBait application.
email	This is the primary email address declared by the user.
user_about_me	This is the user's self description (the "About me" section of their profile).
user_activities	This is a list of OSN activities (e.g. likes, becoming friends with other users, posting, etc.) as listed in the profile of a user.
user_education_history	The education history of the user.
user_hometown	Hometown location of the user.
user_interests	List of declared interests.
user_likes	List of Facebook pages and other pages that the user has liked.
user_location	The current location (city) of the user, as declared by the user.
user_photos	The list of photos that the user has posted or in which s/he has been tagged.
user_relationships	This contains the user's relationship status, significant other and

	family members.
user_relationship_details	This contains the user's relationship interest (commonly appears as "interested in …").
user_religion_politics	The user's declared religious and political beliefs.
user_status	This is a list of user statuses. A user status is a post that does not include links, videos or photos.
user_tagged_places	This is a list of places a user has been tagged at in photos, videos, statuses and links.
user_videos	A list of videos that the user has posted or has been tagged in.
user_groups	A list of groups that the user is a member of.
user_work_history	The user's work history.

Table 3. OSN presence data retrieved from Facebook.

There are a couple of things that we should note here. The first is that for a typical user, quite a few of these fields will be empty. The second is that some of these fields, e.g., location, political views and religious beliefs will directly correspond to some specific privacy attribute. A sample of data, in JSON format, as returned from the Facebook Graph API is presented in Table 14 in Annnex I. Additionally, a sample of Facebook status update data can be found in Table 15, also in Annex I.Finally, the browsing behaviour data that is collected by the DataBait plugin for the pre-pilots is listed in Table 4.

Name	**Description**
Site Unique Visits	Websites (URL) visited by the user
Site Visits	# of times a user visited a web site (URL)
Time Spent Per Site	Time a user spent during one visit. (time opened the URL at his browser)
Images	Images Uploaded/Accessed/Downloaded by the user.
Videos	Videos Uploaded/Accessed/Downloaded by the user.
Actions	Click on specific element at the website.
Text	Text Uploaded/Accessed at the website.
News	Page views of specific news elements.

Table 4. Browsing behaviour data that is collected by the DataBait plugin.

During the pre-pilots data for 170 users have been obtained. Some basic statistics for the collected OSN data of these are shown in the following table:

	Average	**St. Deviation**	**Maximum**
Likes	182.73	223.34	1728
Posts	303.13	521.33	3569
Images	476.06	674.43	3805

In total there are 23,494 distinct pages liked by the users. The maximum number of likes a page has is 23. There are 4 pages that have been liked by 23 users (whereas there are 19910 pages that have been liked by only one user). It is also interesting to note that the distribution of the number of pages that have received a number of likes appears to roughly follow a power law distribution. This is demonstrated in Figure 2.

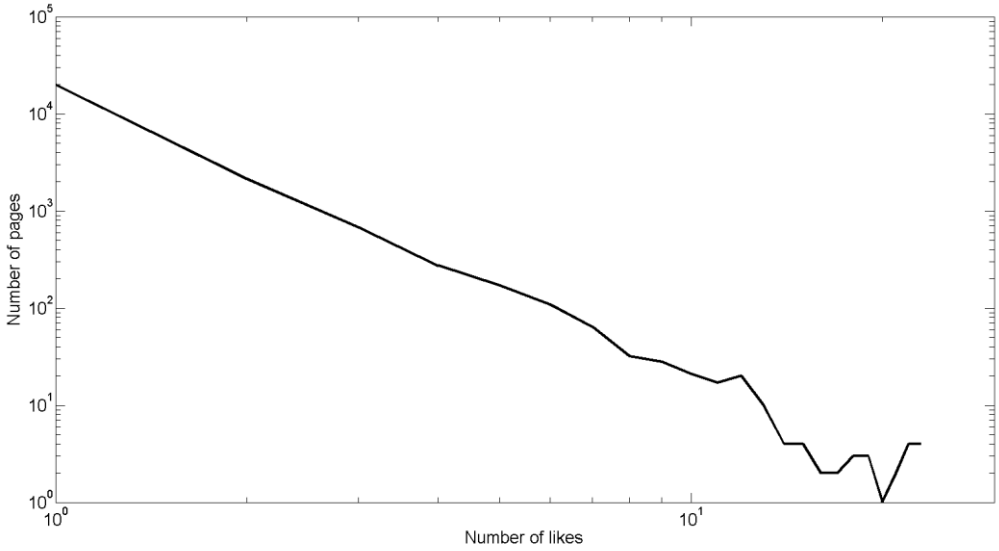

Figure 2. Number of pages that have received a number of likes vs. number of likes. Please note that both axes are in logarithmic scale,

One noteworthy observation on the collected data is that some user classes are not always well represented as others. For instance, Figure 3 shows the distribution of male and female users. Although the majority of participants are males, female participants are relatively well represented, so both classes (male/female) are associated with an adequate number of samples to be used for training.

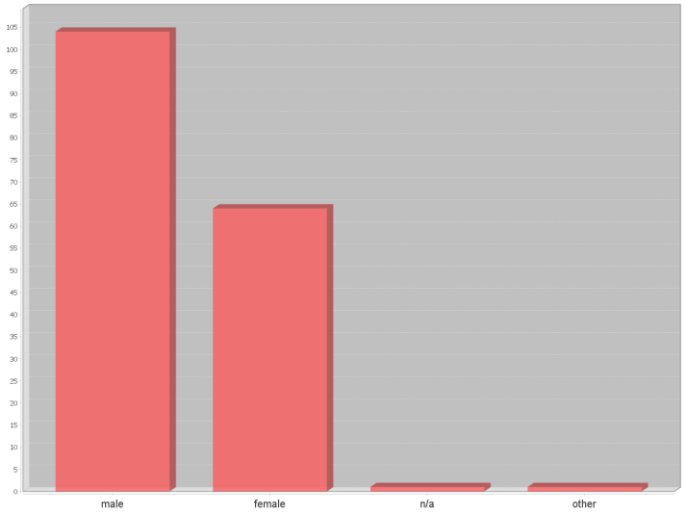

Figure 3. Distribution of male and female users in the pre-pilot

On the other hand, Figure 4 shows the distribution of the participants' sexual orientation. It is clear that homosexual and bisexual participants are a minority in the collected data, and hence it is hard to build accurate classification models for these classes.

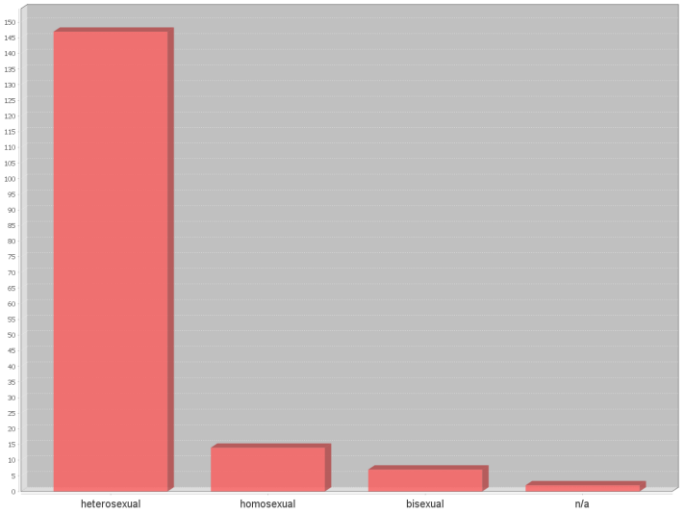

Figure 4. Distribution for the sexuality of users in the pre-pilot

Thorough details about the distributions of the different user classes are included in an online report: http://usemp-mklab.iti.gr/usemp/prepilot_survey_data_statistics.pdf

Dataset usage

Data produced during the pre-pilots and the system operation has been used both for training and for testing of the behavioral and topics-based inference module. More particularly, we utilize the OSN data as predictors for a number of user attributes and use the survey responses as the ground truth. More details can be found in D6.4. One of the main outcomes of this analysis is that different attributes can be predicted with different levels of accuracy.

Legal considerations

The data is used only internally by the consortium members and its usage is governed by the Data License Agreement (DLA) that is also shown through the DataBait interface.

4.USEMP User Profile

Having discussed the data used for the development of the USEMP modules, we proceed to discuss the central data structure of the system during its regular operation, that is, the USEMP user profile. As already mentioned, the USEMP user profile consists of two parts. The first is the set of data retrieved from the user's OSN presence and data from their browsing behavior, whereas the second is the set of USEMP Data Derivatives, mainly the result of the inference algorithms. In the following two sections we proceed to discuss each of them in turn. The term Data Derivatives was taken from the work of Louise Amoore (2011), who links the speculative nature of derivatives in the financial markets with the speculative nature of predictive analytics[11].

4.1. OSN data and browsing behaviour data

The first part of the USEMP user profile that consists of the raw OSN data and browsing behavior data, entails mostly data collected during the pre-pilots as presented in Section 3.2. In particular, the part of the OSN presence data has been presented in Table 3, where the Facebook profile data that is fetched is listed, and the browsing behavior data has been presented in Table 4. We have also listed some basic statistics about the OSN data collected during the pre-pilots. Although these statistics entail a level of uncertainty due to the somewhat limited size of the dataset, they do provide some indication about the expected number of posts, likes and images for the average Facebook user.

Importantly though, there are two additional types of data that may be used during the second iteration of the pilots, but are not used during the pre-pilots. The first is tracking data and the other is OSN data collected from a second OSN.

Web tracking data that will be collected by the DataBait browser plugin related to the user web behaviour are listed in Table 5.

Name	Description
URL of the visited pages	URL of the site visited by a user with Databait plugin installed
URLs within visited pages	URLs of items of interests (videos/images) within a web page
# of Trackers for Site URL	The number of tracking services when a DataBait user visits URL
Tracker	The ID of the tracking services when a DataBait user visits a URL
Tracker email	A Tracker of users email (e.g., google-mail)

Table 5. Trackers data collected by the DataBait plugin.

Regarding data from a second OSN, at the time of writing this deliverable, a final decision about which will be the second OSN that will be integrated to the system has not been made. However, the first candidate is Twitter. Details about the data that can be obtained from Twitter can be found in Annex II.

[11] Louise Amoore, "Data Derivatives On the Emergence of a Security Risk Calculus for Our Times", Theory, Culture & Society 28, nr. 6 (1 november 2011): 24-43.

4.2. USEMP data-derivatives

The raw OSN data is used to perform a number of inferences, each of which is related to some privacy attribute. As a reminder, related attributes are grouped in a set of dimensions. The produced inferences are naturally organized along these dimensions, in a structure that has been presented in D6.1 (the USEMP privacy scoring framework). In the following, we will review the user attributes that we consider as part of the scoring framework, focusing on which modules are used to infer them and with which datasets these modules have been trained. This information is presented for each of the eight privacy dimensions in Table 6 - Table 13. Clearly, many of these dimensions constitute sensitive data in the legal sense, requiring unambiguous explicit consent, as provided in the Data Licensing Agreement.

Attribute	**Module / training dataset**
Age	Personal attribute behavioural detection / MyPersonality Topic-based attribute detection / My Personality
Gender	Personal attribute behavioural detection / MyPersonality Topic-based attribute detection / My Personality
Racial origin	- /YFCC100M
Ethnicity	- / Pre-pilot data
Literacy level	- / Pre-pilot data
Employment status	- / Pre-pilot data
Income level	- / Pre-pilot data
Family status	Personal attribute behavioural detection / MyPersonality Topic-based attribute detection / My Personality

Table 6. Inference modules and training sets for the attributes under the demographics dimension.

Attribute	**Module / training dataset**
Emotional stability	Personal attribute behavioural detection / MyPersonality Topic-based attribute detection / My Personality
Agreeableness	Personal attribute behavioural detection / MyPersonality Topic-based attribute detection / My Personality
Extraversion	Personal attribute behavioural detection / MyPersonality Topic-based attribute detection / My Personality
Conscientiousness	Personal attribute behavioural detection / MyPersonality Topic-based attribute detection / My Personality
Openness	Personal attribute behavioural detection / MyPersonality Topic-based attribute detection / My Personality

Table 7. Inference modules and training sets for the attributes under the psychological traits dimension.

Attribute	**Module / training dataset**
Sexual preference	Personal attribute behavioural detection / MyPersonality Topic-based attribute detection / My Personality

Table 8. Inference modules and training sets for the attributes under the sexual profile dimension.

Attribute	Module / training dataset
Parties	Text similarity / Wikipedia Opinion Mining / SentiWordNet
Political ideology	Personal attribute behavioural detection / MyPersonality Topic based attribute detection / My Personality Text similarity / Wikipedia Opinion Mining / SentiWordNet

Table 9. Inference modules and training sets for the attributes under the political views dimension.

Attribute	Module / training dataset
Supported religion	Personal attribute behavioural detection / MyPersonality Topic-based attribute detection / My Personality Text similarity / Wikipedia Opinion Mining / SentiWordNet

Table 10. Inference modules and training sets for the attributes under the religious beliefs dimension

Attribute	Module / training dataset
Smoking	Large scale visual concept recognition / ImageNet, YFCC100M
Drinking (alcohol)	Large scale visual concept recognition / ImageNet, YFCC100M
Drug use	- / -
Chronic diseases	- / -
Other health factors	- / -

Table 11. Inference modules and training sets for the attributes under the health factors dimension.

Attribute	Module / training dataset
Home	Location estimation module / Location estimation dataset Concept detection / Wikipedia
Work	- / -
Favourited places	- / -
Visited places	Location estimation module / Location estimation dataset Concept detection / Wikipedia

Table 12. Inference modules and training sets for the attributes under the locations dimension.

Attribute	Module / training dataset
Brand attitude	Logo recognition module / Logo recognition dataset, ImageNet
Hobbies	Text similarity / Wikipedia
Devices	Text similarity / Wikipedia

Table 13. Inference modules and training sets for the attributes under the consumer profile dimension.

5.Data Management

We now proceed to discuss a number of additional data issues. In particular, we review the data flows in the USEMP system. Moreover, we examine some storage details, i.e. types of databases to be used and related APIs. This analysis should lead to a user-friendly, intuitive presentation and visualisation of the various data flows and their employment in the context of USEMP research. This information should be available behind the information button on the USEMP platform, constituting critical compliance with the information obligations of the USEMP Consortium Partners. In D3.4 this section will be reframed as a Data Lifecycle Management approach, connecting with the obligations already stipulated in the Data Licensing Agreement and the underlying Personal Data Processing Agreement.

5.1. Data flow and usage

In order to examine how the data flows through the system we will quickly review the USEMP conceptual architecture. The USEMP conceptual architecture is shown in Figure 5Figure 5. Note that this diagram is directly copied from D7.1. Some of the names do not exactly match the names used in this document; any mismatches will be pointed out.

At a high level there are two major parts. The first is referred to as USEMP-TOOLS in the diagram and contains the front-end interface with which the user interacts, as well as components that gather data related to the OSN presence and the web browsing behavior of the user. The second part is referred to as USEMP-SS (USEMP system services) in the diagram and contains all storage and processing components.

The data follows a linear flow through the system. Most data originates from the modules that collect the OSN presence data and the browsing behavior data and is stored in the historical database. The data is then directed to the appropriate data-driven modules (referred to as "Technical components" in the diagram). Subsequently, the overall profile of the user is built and maintained: this is carried out by the disclosure scoring framework based on the results produced by the data-driven modules. Personal data value estimates are also computed. Eventually, the overall user profile is stored in the disclosure database[12] and results are retrieved and shown in the user interface when required. The diagram also shows the external data that are used for training and tuning the data driven modules as also flowing to the system, but this can be considered as an offline data flow.

The flow of data through the USEMP platform is controlled by a number of events. We identify the following events and flows of data:

- A new user registers with the USEMP platform. When a new user registers with DataBait, their textual OSN presence data is fetched and stored in a NoSQL data store resident within the backend servers. Image data is collated independently as a list of URIs and passed to a backend process. This process fetches all the images and places them on a segregated disk array accessible only via the image processing algorithms. As the data is fetched, another backend process, according to the types of data that has been fetched, triggers the modules defined in Table 1. The resulting

[12] Although the original name of this database in D7.1 was simply privacy database, here we call it disclosure database for reasons explained in D6.4

output is stored within the disclosure scoring framework and the perceived privacy database, resident in an additional NoSQL data store.

- The user posts some new content or updates his OSN profile through the DataBait tool. The historical database is updated accordingly and then the data flows through the system up to the perceived privacy database.
- The tool checks periodically (e.g., once per day, or every time the user logs in with DataBait), whether the OSN presence data of the user has changed outside DataBait (e.g., direct interaction via the OSN), triggers the appropriate historical database update operations and analysis procedures and finally updates the privacy database.
- The DataBait tool asks for data from the privacy database. The relevant information is retrieved from the database and is sent to the DataBait tool.

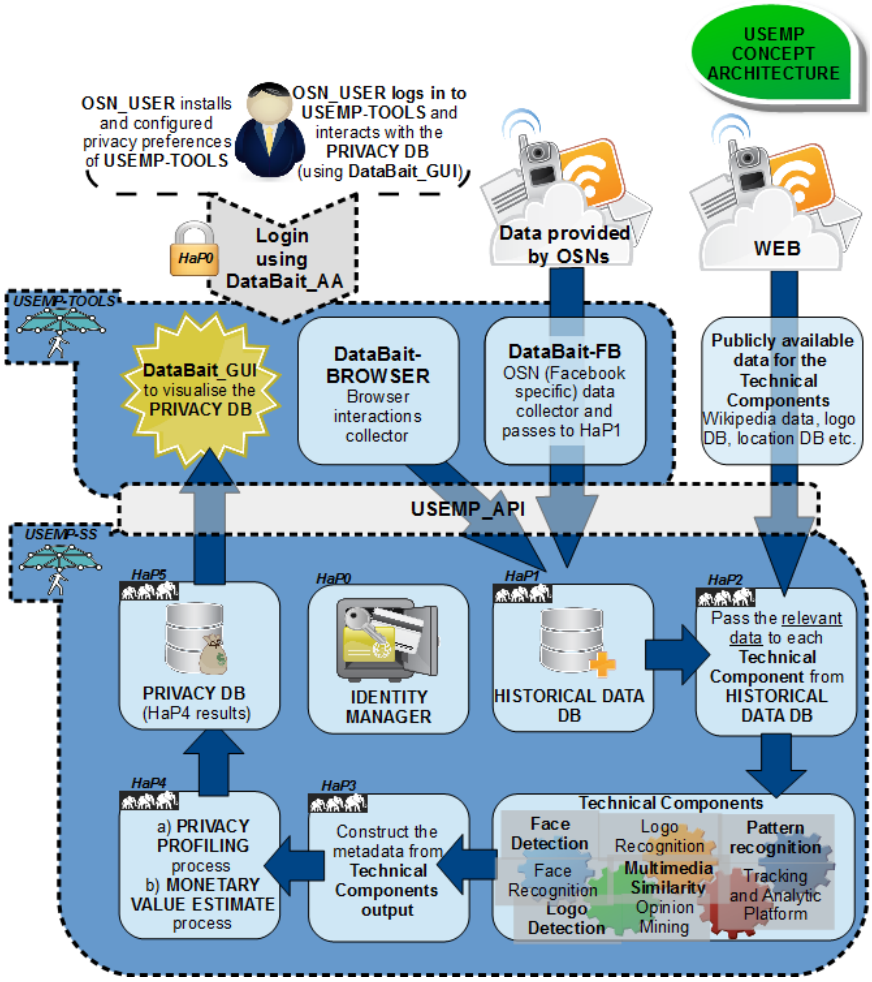

Figure 5. USEMP Conceptual Architecture.

5.2. Storage

USEMP data are stored in a number of segregated databases such that information is resident closest to the elements processing the data. Moreover keeping data segregated means inferences between sets of data can only be performed by those algorithms and processes that have been given specific permission to access the associated data sets.

We have already mentioned the two main databases that are used by the USEMP system: the historical database and the privacy database. Nevertheless, the historical database in

particular consists of three distinct databases and there is an additional database. Each of the databases and storages of the system are presented in more detail in the following.

DataBait identity manager: This is a database used for storing the details of the DataBait users. It contains information relevant to logging into the DataBait service (the User Identity/ Profile), and information relevant to the running of the overall system, and linking between additional identities (e.g., OSN Identities/Profiles). This information is stored in a traditional SQL database within the backend system, accessible only via backend services. The DataBait User Profile consists of the following elements:

- DataBait Username
- DataBait Password Hash (Salted and Hashed)
- DataBait Email
- OSN Tokens
    - Token for each OSN linked to a DataBait account
    - Token for access to the DataBait User Survey
    - Token for access to DataBait from front-end GUI service

Historical database: This contains all OSN presence data as well as the survey data (for those users that have filled the survey). It actually consists of three distinct storages:

a) The survey data, stored within a traditional SQL database, running on the backend, co-located with the survey system. This resides on a system independent of the user profile mentioned above. The survey profile contains a user's answers to all survey questions and associated completion state.

b) A database that stores textual OSN data. Data stored in this database is dependent upon the information provided by the particular OSN and is therefore stored within a NoSQL document store such that it is capable of handling a variety of data types for which no inherent structure is known in advance. Facebook data is the primary constituent of information stored during the pre-pilots, however the system has been designed to be capable of handling any kind of textual OSN data. This data is typically in a structured JSON format, and as such can easily be queried and retrieved within the system.

c) A distributed storage array for OSN imagery data. This imagery data is stored within a distributed disk array within the DataBait backend systems, and is accessible solely from the image processing server hosting the imagery extraction algorithms. Images are stored with the same hierarchy as they are represented within the associated OSN, however attributed textual data is not stored together with them; it remains segregated in the OSN textual data store.

Perceived privacy database: The output of the scoring framework is stored in an additional NoSQL document store, segregated from OSN data and user profile data. This allows for greater flexibility is experimenting with algorithms which may change, while allowing for the primary OSN data store to be accessible in a purely read only state for the purpose of data processing. Moreover it allows for multiple versions of the scoring framework to reside on different backend servers while sharing a single OSN store for data that remain unchanged. The resulting output is in the format specified in D6.1.

6.Summary

This deliverable looked at the data used for the development of USEMP tools and handled during the operation of the USEMP system. Our analysis first identified a number of data requirements organized in two categories. The first is related to a number of modules specified by the functional requirements, whereas the second is about the overall USEMP user profile. We examined each of these in turn and we also had a careful look at the set of external datasets that are used for training and evaluating the USEMP data-driven modules. We also looked at various data management issues, such as the description of the flow of data though the system, as well as storage issues.

Annex I – Internal Facebook representations

In this Annex, samples of actual Facebook data in their original JSON format are listed. To start with, Table 14 provides the basic data for some random Facebook user.


```
{  "status":true,
   "message":"Successful",
   "data":{
      "id":"1424672444497579",
      "metadata":null,
      "type":null,
      "name":"John Smith",
      "firstName":"John",
      "middleName":null,
      "lastName":"Smith",
      "link":"https://www.facebook.com/app_scoped_user_id/1424672444497579/",
      "bio":null,
      "quotes":null,
      "about":null,
      "relationshipStatus":null,
      "religion":null,
      "website":null,
      "birthday":"07/08/1987",
      "email":"johnsmith@hwcomms.com",
      "timezone":1.0,
      "verified":true,
      "gender":"male",
      "political":null,
      "locale":"en_GB",
      "username":null,
      "picture":null,
      "hometown":null,
      "location":null,
      "significantOther":null,
      "updatedTime":1426693552000,
      "thirdPartyId":null,
      "currency":null,
      "tokenForBusiness":null,
      "interestedIn":[  ],
      "meetingFor":[  ],
      "work":[  ],
      "education":[  ],
      "sports":[  ],
      "favoriteTeams":[
         {  "id":"48835794824",
            "metadata":null,
            "type":null,
            "name":"England Rugby"  }
      ],
      "languages":[
         {  "id":"106059522759137",
            "metadata":null,
            "type":null,
            "name":"English" }
      ],
      "birthdayAsDate":552700800000,
      "hometownName":null
   }
}
```


Table 14. Sample of basic Facebook data in JSON format.

Table 15 lists a sample of the status update data for some random Facebook user.

```
{
  "status":true,
  "message":"Retrieved User Statuses.",
  "data":[
    {  "id":"1426831377615019",
       "metadata":null,
       "type":null,
       "name":null,
       "from":{ "id":"1424672444497579",
                "metadata":null,
                "type":null,
                "name":"John Smith"},
       "message":"At the USEMP meeting.",
       "place":{
         "id":"166154786735301",
         "metadata":null,
         "type":null,
         "name":"Kulturens hus Lulea",
         "location":{
           "street":"Skeppsbrogatan 17",
           "city":"Lulea",
           "state":null,
           "country":"Sweden",
           "zip":"971 79",
           "latitude":65.585433066273,
           "longitude":22.151157817905
         },
         "locationAsString":"{\"zip\":\"971 79\",\"street\":\"Skeppsbrogatan
17\",\"longitude\":22.151157817905,\"latitude\":65.585433066273,\"country\":\"Sweden\",\"city\":\"LuleÃ¥\"}",
         "categoryList":[ ]
       },
       "updatedTime":1426671810000,
       "likes":[
         {  "id":"1424672444497579",
           "metadata":null,
           "type":null,
           "name":"John Smith" }
       ],
       "comments":[
         {
           "id":"1426831377615019_1426895020941988",
           "metadata":null,
           "type":null,
           "from":{  "id":"1424672444497579",
                     "metadata":null,
                     "type":null,
                     "name":"John Smith",
                     "category":null },
           "message":"another comment.",
           "createdTime":1426686291000,
           "likes":null,
           "likeCount":0,
           "canRemove":true,
           "userLikes":false,
           "parent":null,
           "comments":null,
           "attachment":null
         }
       ]
    }
  ]
}
```


Table 15. Sample of Facebook status update data as returned from the Facebook API.

Annex II – Twitter data

In this Annex we look at OSN data that may be collected from Twitter. When querying the Twitter API[13] for the details of a user, a large amount of information are returned, including some that may seem unlikely to be linked with personal information, e.g., information about the visual design of the user's profile. The information about a user that is returned by the Twitter API is presented in Table 16. Note that some fields that are of limited interest are only included for completeness and are grouped together in the last entry of the table.

Field	Description
id / id_str	The integer and string representation respectively of the unique identifier for some user in Twitter.
created_at	The date and time that the user account was created on Twitter.
description	A description of the account provided by the user
entities	Entities which have been parsed out of the url or description fields defined by the user.
favourites_count	The number of tweets this user has favorited in the account's lifetime.
followers_count	The number of followers this account currently has.
friends_count	The number of users this account is following.
geo_enabled	When true, indicates that the user has enabled the possibility of geotagging their Tweets.
is_translator	When true, indicates that the user is a participant in Twitter'stranslator community.
lang	A code representing the user's self-declared user interface language.
listed_count	The number of public lists that this user is a member of.
location	The user-defined location for this account's profile.
name	The name of the user, as they've defined it.
notifications	Indicates whether the authenticated user has chosen to receive this user's tweets by SMS
protected	When true, indicates that this user has chosen to protect their Tweets.
screen_name	The screen name, handle, or alias that this user identifies themselves with. screen_names are unique but subject to change.
status	If possible, the user's most recent tweet or retweet.
statuses_count	The number of tweets (including retweets) issued by the user.
time_zone	A string describing the Time Zone this user declares himself within.
url	A URL provided by the user in association with their profile.
utc_offset	The offset from GMT/UTC in seconds.

[13] https://dev.twitter.com/overview/api

verified	When true, indicates that the user has a verified account.
withheld_in_countries	When present, indicates a textual representation of the two-letter country codes this user is withheld from.
withheld_scope	When present, indicates whether the content being withheld is the “status” or a “user.”
Other information	Some additional fields; some are related to various preferences of the user about the appearance of his/her profile, others about the OSN link between the user that issues the query and the user about whom Twitter is queried. Most of this information is of rather limited interest for USEMP purposes (with the exception of images), its use will be more thoroughly evaluated in the future though: • contributors_enabled • default_profile • default_profile_image • follow_request_sent • following • profile_background_color • profile_background_image_url • profile_background_image_url_https • profile_background_tile • profile_banner_url • profile_image_url • profile_image_url_https • profile_link_color • profile_sidebar_border_color • profile_sidebar_fill_color • profile_text_color • profile_use_background_image • show_all_inline_media

Table 16. OSN user data that is returned from the Twitter API.

An instance of data returned from the Twitter API about some user is shown in Table 17.


```
{
	"id":"346293649",
	"name":"John Smith",
	"screenName":"John Smith",
	"location":"",
	"description":"",
	"isContributorsEnabled":"false",
	"profileImageUrl":"http://abs.twimg.com/sticky/default_profile/default_profile_1_normal.png",
	"profileImageUrlHttps":"https://abs.twimg.com/sticky/default_profile/default_profile_1_normal.png",
	"url":"null",
	"isProtected":"false",
	"followersCount":"5",
	"status":"null",
	"profileBackgroundColor":"C0DEED",
	"profileTextColor":"333333",
	"profileLinkColor":"0084B4",
	"profileSidebarFillColor":"DDEEF6",
	"profileSidebarBorderColor":"C0DEED",
	"profileUseBackgroundImage":"true",
	"showAllInlineMedia":"false,
	"friendsCount":"2",
	"createdAt":"Wed Nov 30 15:31:16 EET 2011",
```

```
"favouritesCount":"0",
"utcOffset":"-1",
"timeZone":"null",
"profileBackgroundImageUrl":"http://abs.twimg.com/images/themes/theme1/bg.png",
"profileBackgroundImageUrlHttps":"https://abs.twimg.com/images/themes/theme1/bg.png",
"profileBackgroundTiled":"false",
"lang":"en",
"statusesCount":"0",
"isGeoEnabled":"false",
"isVerified":"false",
"translator":"false",
"listedCount":"0",
"isFollowRequestSent":"false"
}
```


Table 17. Instance of user data returned from the Twitter API in JSON format.

Data posted on Twitter are retrieved separately using the appropriate API calls. Information provided by Twitter regarding posts is listed in Table 18.

Field	**Description**
id / id_str	The integer and string representation respectively of the unique identifier for the tweet.
created_at	UTC time when this tweet was created.
contributors	A set of users (usually only one) that contributed to the authorship of the tweet, on behalf of the official tweet author.
coordinates	Represents the geographic location of this tweet as reported by the user or client application.
entities	Entities which have been parsed out of the text of the tweet.
favorite_count	Indicates approximately how many times this tweet has been “favorited”.
in_reply_to_screen_name	If the tweet is a reply, this field will contain the screen name of the original tweet’s author.
in_reply_to_status_id	If the tweet is a reply, this field will contain the integer representation of the original tweet’s id.
in_reply_to_user_id	If the tweet is a reply, this field will contain the integer representation of the original tweet’s author id
lang	Language identifier for the machine-detected language of the tweet text.
place	When present, indicates that the tweet is associated (but not necessarily originating from) a place.
possibly_sensitive	This field only surfaces when a tweet contains a link. The meaning of the field doesn’t pertain to the tweet content itself, but it is an indicator that the URL contained in the tweet may contain content or media identified as sensitive content. Sensitivity of content is decided based on flags provided by the poster.
scopes	A set of key-value pairs indicating the intended contextual delivery of the containing tweet. Currently used by Twitter’s promoted products.
retweet_count	Number of times this tweet has been retweeted.
retweeted_status	If the tweet is a retweet, this field contains a representation of

	the original tweet.
source	Utility used to post the tweet, as an HTML-formatted string. Tweets from the Twitter website have a source value of web.
Text	The actual text of the tweet.
Truncated	Indicates whether the value of the text parameter was truncated due to its length,
User	The user who posted this tweet.
withheld_copyright	When present and set to “true”, it indicates that this piece of content has been withheld due to a DMCA complaint.
withheld_in_countries	When present, indicates a list of uppercase two-letter country codes this content is withheld from.
withheld_scope	When present, indicates whether the content being withheld is the “status” or a “user.”

Table 18. Fields about a tweet returned by the Twitter API.

Finally, a sample of a tweet as returned by the Twitter API in JSON format can be found in Table 19.


```
{
  "filter_level":"low",
  "retweeted":false,
  "in_reply_to_screen_name":null,
  "possibly_sensitive":false,
  "truncated":false,
  "lang":"en",
  "in_reply_to_status_id_str":null,
  "id":583981737207844864,
  "in_reply_to_user_id_str":null,
  "timestamp_ms":"1428067070663",
  "in_reply_to_status_id":null,
  "created_at":"Fri Apr 03 13:17:50 +0000 2015",
  "favorite_count":0,
  "place":null,
  "coordinates":null,
  "text":"Town v Cherries our main feature from 4 #itfc. Also talking @ipswichspeedway, @MildenhallFT, @IpswichHoops
@PorscheRaces, @S_Nat_Bangers.",
  "contributors":null,
  "geo":null,
  "entities":{
    "trends":[
    ],
    "symbols":[ ],
    "urls":[ ],
    "hashtags":[
      {
        "text":"itfc",
        "indices":[ 40, 45 ]
      }
    ],
    "user_mentions":[
      {
        "id":237654058,
        "name":"Spedeworth",
        "indices":[ 122,  136 ],
        "screen_name":"S_Nat_Bangers",
```

```
            "id_str":"237654058"
         }
      ]
   },
   "source":"<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client<\/a>",
   "favorited":false,
   "in_reply_to_user_id":null,
   "retweet_count":0,
   "id_str":"583981737207844864",
   "user":{
      "location":"Suffolk, UK",
      "default_profile":true,
      "statuses_count":2690,
      "profile_background_tile":false,
      "lang":"en",
      "profile_link_color":"0084B4",
      "id":21892954,
      "following":null,
      "favourites_count":0,
      "protected":false,
      "profile_text_color":"333333",
      "verified":false,
      "description":null,
      "contributors_enabled":false,
      "profile_sidebar_border_color":"C0DEED",
      "name":"BBC Suffolk Sport",
      "profile_background_color":"C0DEED",
      "created_at":"Wed Feb 25 17:44:35 +0000 2009",
      "default_profile_image":false,
      "followers_count":2894,
      "profile_image_url_https":"https://pbs.twimg.com/profile/80bc_suffolk_logo_norm.jpg",
      "geo_enabled":false,
      "profile_background_image_url":"http://abs.twimg.com/images/themes/theme1/bg.png",
      "profile_background_image_url_https":"https://abs.twimg.com/images/themes/theme1/bg.png",
      "follow_request_sent":null,
      "url":"http://www.bbc.co.uk/suffolk/sport",
      "utc_offset":3600,
      "time_zone":"London",
      "notifications":null,
      "profile_use_background_image":true,
      "friends_count":280,
      "profile_sidebar_fill_color":"DDEEF6",
      "screen_name":"bbcsuffolksport",
      "id_str":"21892954",
      "profile_image_url":"http://pbs.twimg.com/profile_images/83410906/bbc_suffolk_2009_logo_203x152_normal.jpg",
      "listed_count":62,
      "is_translator":false
   }
}
```


Table 19. Sample of tweet as returned from the Twitter API.

Bibliography

L. Amoore (2011) Data Derivatives on the Emergence of a Security Risk Calculus for Our Times, Theory, Culture & Society 28(6), pp. 24-43.

Art. 29 Working Party, Opinion 5/2014 on anonymisation techniques (WP216)

S. Baccianella, A. Esuli, F. Sebastiani (2010) SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining

D. Bouamor, A. Popescu, N. Semmar, P. Zweigenbaum (2013) Building Specialized Bilingual Lexicons Using Large-Scale Background Knowledge. Proc. of *EMNLP 2013,* Seattle, USA.

J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei. (2009). ImageNet: A Large-Scale Hierarchical Image Database. Proceedings of CVPR 2009.

B. Ionescu, A. Popescu, M. Lupu, A. Gınsca, H. Müller. (2014). Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In MediaEval 2014 Workshop, Barcelona, Spain.

Y. Jia. (2013). Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding. http://caffe.berkeleyvision.org

M. Marchand, A. L. Ginsca, R. Besançon, O. Mesnard (2013) [LVIC-LIMSI]: Using Syntactic Features and Multi-polarity Words for Sentiment Analysis in Twitter. Working notes of SemEval 2013.

P. Ohm, “Broken Promises of Privacy: Responding tothe Surprising Failure of Anonymization”, UCLA Law Review 57 (2010): 1701–77.

S. Papadopoulos, D. Corney, L. Aiello. (2014). SNOW 2014 Data Challenge: Assessing the Performance of News Topic Detection Methods in Social Media. Proceedings of the SNOW 2014 Data Challenge co-located with (WWW 2014), pp. 1-8

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei. (2014). ImageNet Large Scale Visual Recognition Challenge. arXiv technical report: http://arxiv.org/abs/1409.0575

S. Romberg, L. Garcia Pueyo, R. Lienhart, R. van Zwol. (2011). Scalable Logo Recognition in Real-World Images. Proceedings of ACM International Conference on Multimedia Retrieval 2011 (ICMR11), Trento

P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun. (2013). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. arXiv technical report: http://arxiv.org/abs/1312.6229

B. Schneier. A Taxonomy of Social Networking Data, Security & Privacy, IEEE , vol.8, no.4, pp.88,88, July-Aug. 2010

B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L.-J. Li (2015) The New Data and New Challenges in Multimedia Research. Arxiv preprint http://arxiv.org/abs/1503.01817 (consulted on 12/03/2015)

S. Zerr, S. Siersdorfer, J. Hare, E. Demidova. I Know What You Did Last Summer!: Privacy-Aware Image Classification and Search. SIGIR 2