Knowledge Discovery in distributed Social Web sharing activities

Simon Scerri, Keith Cortis, Ismael Rivera, Siegfried Handschuh
Digital Enterprise Research Institute
National University of Ireland, Galway
firstname.lastname@deri.org

ABSTRACT

Taking into consideration the steady shift towards information digitisation, an increasing number of approaches are targeting the unification of the user's digital "Personal Information Sphere" to increase user awareness, provide single-point management, and enable context-driven recommendation. The Personal Information Sphere refers to both conventional information such as semi/structured information on the user's personal devices and online accounts, but also in the form of more abstract personal information such as a user's presence and activities. Online activities constitute a rich source for mining this type of personal information, since they are usually the only means by which a typical user consciously puts effort into sharing their activities. In view of this opportunity, we present an approach to extract implicit presence knowledge embedded in multiple streams of heterogeneous online posts. Semantic Web technologies are applied on top of syntactic analysis to extract and map entities onto a personal knowledge base, itself integrated within the wider context of the Semantic Web. For the purpose, we introduce the DLPO ontology—a concise ontology that captures all facets of dynamic personal information shared through online posts, as well their various derived links to personal and global semantic data clouds. Based on this conceptualisation, we outline the information extraction techniques targeted by our approach and present an as yet theoretical use-case to substantiate it.

Keywords

Social Web, Microposts, Presence, Ontologies, Personal Information Management

1. INTRODUCTION

The di.me project[1] is one of many initiatives targeting the unification of a user's personal information sphere across heterogeneous sources, with the aim of providing an intelligent and proactive system—the di.me userware—that assists the user in their day-to-day activities.

The di.me userware generates and constantly updates a representation of the user's Personal Information Model (PIM), based on a comprehensive modelling framework that combines various ontologies provided by the OSCA Foundation (OSCAF)[2]. Thus, the PIM serves as an integrated personal Knowledge Base (KB) containing all available knowledge about the user (i.e. their devices, accounts, social relationships, resources, activities, etc.), as crawled and mined by the di.me userware. It is however not a self-contained knowledge model, also containing references to resources in open repositories. Knowledge stored in the PIM enables advanced features such as distributed personal information management, improved search and retrieval, context-awareness and context-dependant recommendation.

In the context of di.me, the term personal information sphere refers to not just conventional structured or semi-structured data (e.g. files, folder structures, contact lists, photo albums, status messages, etc.), as was the case in earlier initiatives such as the Social Semantic Desktop [22]. Di.me also covers unstructured, more abstract forms of personal information, including complex concepts such as user contexts, situations, physical and online presence. In order to elicit this type of personal information, di.me identifies two types of sources: sensors (presence information relayed by device-embedded sensors, user attention monitoring, etc.) and social sharing activities (serving as 'virtual sensors' [5]).

In this paper, we focus on the latter, as a novel and rich source for enriching the PIM with presence-related knowledge. Thus, the main motivation for this work is the extraction of information from multiple streams of heterogeneous online-post data[3], by both the users and their contacts, in order to generate valuable outputs. More specifically, we exploit online sharing activities in order to i) enrich the PIM with discovered personal and social knowledge (e.g. detecting a user's current activity, availability, learning who is in the same area, doing a similar activity, discussing the same topic, etc.), ii) semantically link post items across personal social networks (e.g. Facebook, Twitter, LinkedIn posts

[1] http://dime-project.eu/

[2] http://www.oscaf.org/ – OSCAF ontologies have been adopted by various initiatives, including di.me.

[3] Although at the conceptual level we consider all types and lengths of online posts, the technical approach is more focused on the shorter, so-called microposts.

Published as part of the #MSM2012 Workshop proceedings, available online as CEUR Vol-838, at: http://ceur-ws.org/Vol-838
#MSM2012, April 16, 2012, Lyon, France.

about the same topic, event, location, video, etc.), and iii) enable social-based recommendations (e.g. the user is shown contacts that have similar interests, are in the same area, are discussing similar topics, doing related activities, etc.).

In line with the above objectives, this paper provides two research contributions: the provision of a suitable model for the representation of online posts and their embedded knowledge, and the design of comprehensive semantic lifting techniques for the extraction of the knowledge itself. The first contribution consists of the LivePost Ontology (DLPO), an open knowledge representation standard[4] integrated within the existing PIM models, that provides rich conceptualisations of various types of online posts, covering more emerging Social Web sharing features than any of the available standards.

The second contribution is an overview of the techniques to be employed for the extraction of semantics from microposts. Online post items contain both easily-acquired semi/structured data (e.g. hyperlinks, creator, date and time, people tagged, nearby places, etc.) as well as hidden abstract data that requires the application of advanced linguistic techniques. The target of the semantic lifting process is to break down retrieved posts into one or more specific subtypes (e.g. message post, image post, checkin, etc.), and enrich them with clear semantics (including links to existing PIM and/or Semantic Web resources). The Information Extraction (IE) techniques employed include shallow parsing techniques such as Named-Entity Extraction (NEE), keyword extraction and hyperlink resolution, followed by more sophisticated analysis such as Named-Entity Resolution (NER) and co-reference resolution, topic extraction, and time window analysis.

After comparing related work (Section 2) and outlining the approach (Section 3), we provide examples of how our approach can result in a semantic representation that is integrated within the PIM (Section 4), before providing some concluding remarks and directions for future work (Section 5).

2. RELATED WORK

In the light of our requirements, we here outline and compare related approaches. The use of social data (in the form of microposts) as input data for providing some form of recommendations is not an entirely novel concept, and has been in fact applied by a number of other approaches [7], [6], [8], [1], [26]. In particular, Chang and Sun [8] analyse a dataset of Point of Interests (POIs) collected from Facebook Places to construct a prediction model for a user's future locations. The BOTTARI mobile app [7] exploits social context to provide items to the user, relevant to their location. Our approach is more similar to the latter, collecting information about a user's and their contacts' presence, such as to enable the discovery of information which would otherwise easily be missed, e.g., contacts discussing the same topics, travelling to the same city, etc. The collected information will also be used to provide context-aware suggestions (nearby POIs that are recommended by trusted contacts) and warnings (untrusted persons in the vicinity).

[4] `http://www.semanticdesktop.org/ontologies/dlpo/`— a candidate OSCAF submission

Given our emphasis on knowledge representation, Table 1 compares existing vocabularies (or a combination of) modelling the required domains—User Presence and Social Web sharing through posts—against the DLPO. As is evident, most approaches are targeted towards one domain, with only a few supporting cross-domain modelling. The Semantically-Interlinked Online Communities (SIOC) Ontology for instance, is more oriented towards Social Web sharing, as its original aim was to interlink online communities [3]. Although it caters for online posts (denoted by a full circle), it does not attempt to make any sort of link between user posts and user presence, which is a missed opportunity given that a large number of microposts are linked to physical and online user activities/experiences[5] (i.e. user presence[6]). A proposed combination of the SIOC(T), Friend of a Friend (FOAF) [4] and Simple Knowledge Organization System (SKOS) [17] ontologies[7] is also unable to represent any form of user presence in relation to the posts. The Bottari Ontology [7] is another SIOC extension which supports relationships between posts, locations and user sentiment, as extracted from tweets. Another vocabulary that provides for representations of online posts in the context of user presence is the Online Presence Ontology (OPO) [23], which models a user's current activities on online services. But since the OPO does not cover all physical presence aspects, it is insufficient for our representation needs. The PreSense Ontology [5] reuses OPO vocabulary to effectively cater for the representation of both physical and online presence. It also makes a connection between a user's presence and their online status streams, which can serve as 'virtual' presence sensors. However, without the re-use of additional OPO vocabulary, PreSense remains unable to provide the comprehensive modelling of online posts that we need.

Another requirement is to be able to decompose a post into multiple concurrent sub-types (e.g. into a status message, an image photo upload, and a check-in). Post decomposition has the advantage of improving retrieval (e.g. user later looks for all items posted in an area, thus showing only the meta-data of specific posts) and deletion (e.g. it's much easier to remove any posts related to a deleted PIM concept, since any related PIM concepts are directly linked to their corresponding posts, unlike in SIOCT where relations between a PIM concept and a resource are not directly known). Additionally, vocabularies such as the Privacy Preference Ontology (PPO) [21] could be employed to enable a user to restrict or allow access to only some types of subposts (e.g. share all types of posts with a contact group, except for place check-ins). Although most approaches in Table 1 provide for multiple post types (denoted by a half-full circle), the DLPO provides the best representation of concurrent posts. The original SIOC vocabulary did not even distinguish between different kinds of posts, a limitation which

[5] 77.7% of tweets consist of 'conversations' and 'pointless babble', both of which are considered a source of user presence information: `http://mashable.com/2009/08/12/twitter-analysis/`

[6] By the term 'presence' we refer to both online (activities e.g. check-ins, posts, liking; interactions e.g. playing a game, chatting; availability, visibility, etc.) and physical (activities e.g. travelling, walking, working; current location, nearby people and places, etc.) user experiences

[7] `http://sioc-project.org/node/341`

Table 1: Knowledge Models

	SIOC	SIOCT	SIOC(T)+FOAF+SKOS	Bottari	OPO	PreSense	DLPO
Online Posts (General)	●	●	●	●	●	○	●
Post Multi/Sub-typing	○	◐	◐	◐	◐	○	●
Microposts	○	●	●	●	●	●	●
Online Presence	○	○	○	○	●	●	●
Physical Presence	○	○	○	◐	◐	●	●
Online Sharing Practices	◐	◐	◐	◐	○	○	●

was addressed by a later extension (SIOCT[8]) that provides additional sub-types such as *sioc:MicroBlogPost*.

A last quality on which we base our comparison is the modelling support for emerging Social Web sharing practices and interactions such as: post item replies, resharing (or retweeting), endorsement ('liking', starring, favouriting), and general time-awareness (i.e. timestamps, succession of posts, etc.). Neither of the surveyed vocabularies provide for all of the above features, in contrast to the DLPO. At this stage it is necessary to point out that the DLPO does not ignore existing established vocabularies such as SIOC, and in fact, fully re-uses some of its elements. Apart from providing all qualities listed in Table 1, the DLPO stands out for another unmatched characteristic—it is integrated within an entire framework of ontologies targeting the representation of a user's entire personal information sphere. Thus, instances of the DLPO automatically become part of, and are tightly integrated within, a representation for the user's entire PIM.

After considering alternate knowledge models, we now compare the di.me approach to related approaches, against the IE and semantic lifting techniques they employ (Table 2). Some form of linguistic analysis (keyword/topic extraction, NEE for various entity types) is performed by all. The tools provided by [15] and [20], present improved NEE techniques based on informal communication—noisy, informal and insufficient information—of microposts such as tweets. On the other hand, NER is only performed by Zoltan and Johann [26], in their approach to construct user profiles from knowledge extracted from microposts. In addition, extracted entities are linked to specific concepts within the Linked Open Data (LOD)[9] Cloud. This constitutes what we refer to as 'semantic lifting', whereby structured/unstructured data is lifted onto standard knowledge models such as ontologies. However, our approach is unique because, apart from community KBs (such as LOD), we also enrich a personal KB—the user's PIM. This has the obvious advantage that it consists only of personal data, making it easier to determine equivalence between entities in microposts and PIM items. In addition to the generic techniques listed in Table 2, we also employ techniques such as hyperlink resolution and time window analysis.

3. APPROACH

In this section we present both the conceptual and the practical aspects of our approach. The former consists of an exercise in knowledge modelling, with the resulting ontology then serving as a standard for data integration across multiple and heterogeneous online data sources. The practical side of our approach targets the semantic lifting of data from these sources, including straightforward lifting of structured data, before moving on to the more challenging extraction of semantics from semi-structured and unstructured data.

3.1 MODELLING ONLINE POSTS

The motivation for our approach outlines four major requirements for this modelling task:

i) to support and re-use existing standards—particularly the W3C submission for SIOC [3]

ii) integration within established PIM knowledge models

iii) semantic decomposition into independent sub-posts

iv) representation of emerging online Social Web practices

The DLPO, of which an overview is given by Fig. 1, satisfies all of the above requirements. To adhere to the first, the superclass *dlpo:LivePost* is itself an extension of the generic *sioc:Post*, inheriting all SIOC properties that apply (e.g. sioc:has_creator, sioc:hasTopic). DLPO also introduces two subproperties of the core SIOC properties *sioc:has_reply* and *sioc:reply_Of*, for use within the DLPO context: *dlpo:hasReply*, *dlpo:replyOf*.

The second requirement ensures that the information represented by DLPO is firmly integrated within the wider context of distributed personal information modelling. For the purpose, embracing the Information Element (NIE) ontology, *dlpo:LivePost* is also a subclass of *nie:InformationElement*. This means that a post instance will inherit all properties that apply, e.g. *nie:created* and its DLPO extension *dlpo:timestamp*, both of which are subproperties of *dcterms:created*. The purpose of the NIE ontologies[10] is to provide a vocabulary for describing information elements which are commonly present on a source hosting information belonging to a user. The Personal Information Model Ontology (PIMO)[11] is then used to generate a representation of the user's mental model, by abstracting multiple information element occurrences (e.g. different contacts for a person, formats for a document, etc.) onto a unique and integrated model. Sources for populating the PIM include personal devices and online accounts. The Account Ontology (DAO) enables the representation of online sources (e.g. Facebook, LinkedIn), such that personal information elicited from within can retain the link to

[8] http://rdfs.org/sioc/types
[9] http://richard.cyganiak.de/2007/10/lod/
[10] http://www.semanticdesktop.org/ontologies/nie/
[11] http://www.semanticdesktop.org/ontologies/2007/11/01/pimo/

Table 2: General Approaches

	[7]	[24]	[6]	[9]	[8]	[1]	[26]	[18]	[15]	[20]	di.me
Keyword Extraction	○	○	●	●	○	○	●	○	○	○	●
Topic Extraction	○	○	●	○	●	●	●	●	○	●	●
NEE (Events)	○	◐	○	●	◐	●	○	○	○	◐	●
NEE (People)	○	◐	●	●	●	●	●	○	●	●	●
NEE (Activities)	○	◐	○	○	◐	○	○	○	○	○	●
NEE (Locations)	◐	◐	●	○	●	●	●	●	●	●	●
NER	○	○	○	○	○	○	●	○	○	○	●
Semantic Lifting	○	○	○	○	○	○	●	○	○	○	●

each source. *dao:Account* is a subclass of *sioc:Container*, and *dao:source* a subproperty of *sioc:hasContainer*. Another relevant OSCAF ontology is the Annotation Ontology (NAO)[12], which provides a vocabulary for defining generic, domain-independent relationships between related resources. In the DLPO, the NAO is applied to define the online post's creator *nao:creator*, a defining image or symbol *nao:prefSymbol*, topics *nao:hasTopic*, labels *nao:prefLabel*, tags *nao:hasTag*, descriptions *nao:description* and pointers to the post's unique identifier on the source account *nao:externalIdentifier*. The DLPO also extends two NAO properties, the *nao:isRelated* with *dlpo:relatedResource*, and *nao:hasSuperResource* with *dlpo:definingResource*, to create two types of generic relationships between online posts, or their sub-types, to items in the PIM. *dlpo:relatedResource* can be used to create a semantic link between a livepost and the PIM items which it is about (e.g. a post about people, topics, images, events, places, etc. that are known and represented in the PIM). The *dlpo:definingResource* goes one step further, defining a direct relationship between a post subtype and a PIM item (e.g. linking an EventPost to the actual Event which it describes).

Sub-typing online posts is related to the third requirement, that of decomposing a post into semantically distinguishable subposts. Online service accounts commonly distinguish between different types of posts. The DLPO supports these distinctions, and also the fact that post sub-types can either occur individually or, also in conjunction (e.g. a composite Status Message that contains text, a photo, a nearby location, people tagged, etc.). By definition, a *dlpo:LivePost* may consist of multiple subposts, which relationship can be represented through the use of the *dlpo:isComposedOf* property (a subproperty of *nao:hasSubResource*). The DLPO differs between four categories of sub-posts:

1. **Message** - Represents the text-based subpart of online posts. If a message is not in-reply to a previous message (denoted by *dlpo:replyOf*) then it is of subtype *dlpo:Status*, otherwise it is of subtype *dlpo:Comment*.
2. **MultimediaPost** - Represents subposts containing links to multimedia items that are either available online, or that have been uploaded to the online account. This category of subposts is further refined into *dlpo:VideoPost*, *dlpo:ImagePost* and *dlpo:AudioPost*.
3. **WebDocumentPost** - Represents subposts containing links to online text-based containers. Examples range from a note (*dlpo:NotePost*) or blog entry (*dlpo:BlogPost* as a subclass of *sioct:BlogPost*) to other unresolved non-multimedia links (e.g. online article, web pages, etc.).
4. **PresencePost** - Represents subposts relating to a user's presence. This can refer to not only online presence (*dlpo:AvailabilityPost*) but also physical (*dlpo:ActivityPost*, *dlpo:EventPost* or *dlpo:Checkin*).

The use of the *dlpo:definingResource* property is crucial to achieve the required integration of DLPO instance within

[12] `http://www.semanticdesktop.org/ontologies/nao/`

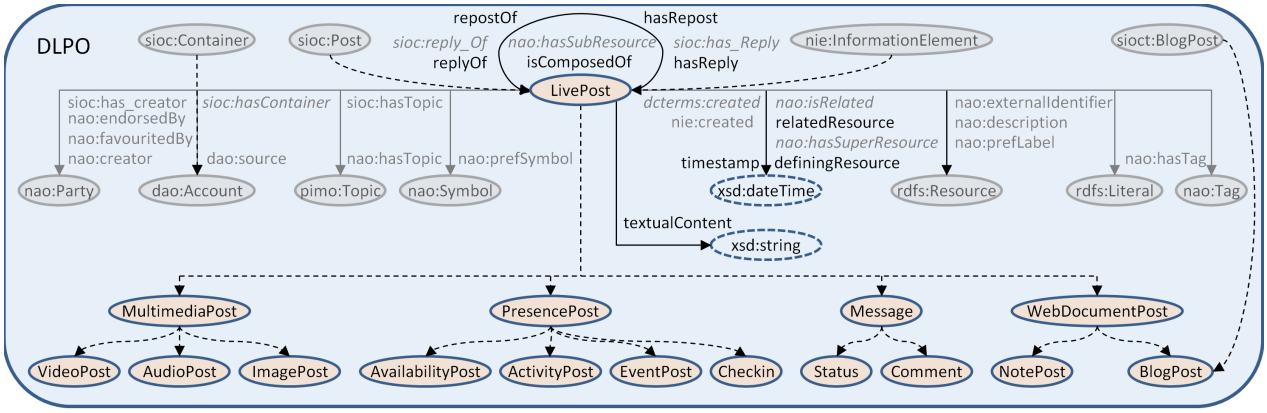

Figure 1: The LivePost Ontology

the PIM. In fact, subtypes are only defined for an online post if it has been determined that it is describing existing PIM items. These items have clearly-defined semantics, and therefore specific subposts will always have specific items as their defining resource, as defined by one of the OSCAF domain ontologies. For example, *dlpo:EventPost* will be related to an *ncal:Event* instance from the Calendar Ontology (NCAL)[13], a *dlpo:Checkin* to a *pimo:Location*, a *dlpo:Video/Image/AudioPost* to a respective instance from the File Ontology (NFO)[14], and an Availability/ActivityPost to one of the available instances provided by the Presence Ontology[15] (DPO) e.g. busy, available, etc. and travelling, working, sleeping, etc.

The fourth modelling requirement is tackled by the introduction and/or adoption of a number of useful properties. These currently include the *dlpo:repostOf* and *dlpo:hasRepost* pair of properties to represent the functionality provided by many online accounts to reshare a (personal or a contact's) post with different social accounts. Additionally, we also reuse the *nao:endorsedBy* and its subproperty *nao:favouritedBy* to represent Social Web functions of 'liking' and favouriting/starring online posts, respectively.

3.2 SEMANTIC LIFTING

This section first explains the transformation from semi-structured posts to a DLPO instance (Sect. 3.2.1), before detailing the syntactic and semantic analysis performed to extract richer information from both semi-structured post metadata and unstructured text. The analysis is decomposed into several tasks, and grouped into one or more semantic annotation pipelines defined using the General Architecture for Text Engineering (GATE) [11] and the ANNIE IE system [10], which includes a variety of algorithms for sentence splitting, gazetteer lookup, etc.

3.2.1 Lifting Semi-Structured data onto the PIM

This phase focuses on the integration of extracted data within the PIM. At this stage, only explicit micropost knowledge is captured and transformed into a DLPO instance. The required XML to RDF[16] transformation is carried out through XSPARQL[17]. The latter combines XQuery[18] and SPARQL[19], thus also enabling the use of external SPARQL endpoints to further enrich the results (e.g. to resolve a location-name based on its coordinates).

Listing 1 is an excerpt of the XSPARQL query created for transforming tweets into *dlpo:StatusMessage* instances. At a glance, it traverses status messages (tweets) contained in an XML document. For each, a `construct` clause creates a new *dlpo:StatusMessage*, and populates it with metadata. The tweet author is created in the nested `construct` clause, and linked to the status message using the *nao:creator* property.

[13] `http://www.semanticdesktop.org/ontologies/ncal/`
[14] `http://www.semanticdesktop.org/ontologies/nfo/`
[15] `http://www.semanticdesktop.org/ontologies/dlpo/`—candidate OSCAF vocabulary for the representation of recurring (online and physical) user presence components.

[16] http://www.w3.org/RDF/
[17] http://xsparql.deri.org/
[18] http://www.w3.org/TR/xquery/
[19] http://www.w3.org/TR/sparql11-query/

```
prefix xsd : <http://www.w3.org/2001/XMLSchema#>
prefix ...

let $doc := "<xml><statuses>...</statuses></xml>"
let $statuses := $doc/statuses/status

return
  for $status in $statuses
    let $id := $status/id
    let $time := $status/created_at
    let $text := $status/text
    let $user := $status/user
    construct
    { _:stm{data($id)} a dlpo:StatusMessage;
      nao:externalIdentifier {data($id)};
      dlpo:timestamp {data($time)}^^xsd:dateTime;
      dlpo:textualContent {data($text)};
      nao:creator _:c{data($id)}.
      {   let $userId := $user/id
        let $name := $user/name
        let $photoUrl := $user/profile_image_url
        let $description := $user/description
        construct {
          _:c{data($id)} a nco:PersonContact;
            nao:externalIdentifier {data($userId)};
            nco:photo {data($photoUrl)};
            nao:description {data($description)};
            nco:fullname {data($name)}. }}}
```

Listing 1: XSPARQL query for Twitter

3.2.2 Preprocessing Unstructured Data

To maximize the quality of results of micropost analysis, our pipeline starts off with the following operations: emoticons removal, abbreviations substitution/removal (using `noslang.com` as an abbreviations dictionary), part-of-speech (POS) tagging, stop words removal and stemming. These tasks execute in that specific order, since e.g. stop words are necessary for our POS tagger, even though they are usually considered as noise by IE algorithms.

3.2.3 Keyword extraction

The majority of research on keyword/keyphrase extraction concentrates on large collections of formal documents (e.g. research papers, news articles), using techniques requiring domain-specific training. Micropost keyword extraction is more challenging due to various reasons; citing their length, the informality of the language, and the higher diversity of topics as examples. Two algorithms suitable for short text (e.g. tweets) are the TF-IDF (term frequency–inverse document frequency) and TextRank [16]. For our approach, we selected TextRank due to i) a better performance compared to classic TF-IDF [25], ii) the use of POS annotations generated by the pipeline to double the accuracy of the output [14], and iii) a design which may reduce the overhead of dynamically adding new documents for the analysis, since convergence is usually achieved after fewer iterations.

3.2.4 Topic extraction

In our context, topics are one level up in abstraction in comparison to keywords. When extracting topics for a resource, they need not explicitly be in the text. Although they may overlap, a topic is also distinct from a category. Whereas the latter are meant for structuring items under a more-or-less fixed taxonomy, the former are intended to function as high-level markers (or tags). In fact, the envisaged di.me userware will allow the user to extend a pre-defined set of topics (instances of *pimo:Topic*).

Techniques for topic discovery from corpora include Latent Dirichlet Allocation (LDA) [2] and Latent Semantic Indexing (LSI) [12]. Both extract topics by clustering words or keyphrases found within. These methods share one important difficulty: finding a meaningful label for the topics. Topic labels are a requirement for the di.me userware—they need to make sense to the user. To overcome this issue, our approach is to involve the user for assigning explicit topics to items in their PIM (including microposts), and recommend topic candidates by finding frequent keywords that co-occur using FP-Growth [13], an algorithm for keyword pattern mining. This algorithm was selected because it offers good performance with limited amount of data, takes minimum support as an argument, leaving the possibility to change how it behaves as the data grows; and is currently one of the fastest approaches to frequent item set mining.

3.2.5 Named-entity extraction

Our NEE task is special since extracted entities are to be mapped onto the user's PIM, which can be seen as a personalised set of entities. Core concepts of the PIMO ontology are also very similar to the generic entities typically extracted by named-entity recognition algorithms (e.g. people, organisations, locations), but they also include more personal (or group) entities (e.g. projects, events, tasks). After extraction, entities will be matched against resources in the PIM, and if that doesn't return any, against the LOD cloud. The matching is syntax-based, comparing named-entity and resource (*rdfs:label*) labels. If several matches are returned, a confidence value will be calculated for each resource, based on keywords extracted from microposts and from descriptive metadata about the resource (e.g. *rdfs:comment*).

NEE taggers based on gazetteers such as [19], [15] or [11] are a good fit for entity extraction where a personal KB may feedback the algorithm with new entities created either directly by the user, or as the result of integrating data from an external KB.

3.2.6 Named-entity & co-reference resolution

To determine orthographic co-reference between terms in natural language text, we use the *orthomatcher* module in GATE. An existing hand-crafted set of rules will be extended for additional entity types handled for di.me, e.g. *pimo:Project*. Due to the short lengths of text, co-reference resolution in microposts is hard. However, grouping them into 'conversations' through (e.g. replies and retweets), the algorithms will have more contextual information to work with. Until entity co-reference is combined with resolution, microposts are not able to enrich the PIM and vice-versa. The fact that "John Doe" and "Mr. Doe" are in fact the same person, is not directly exploitable in semantic lifting. Here, we employ the *Large KB Gazetteer* GATE module to query the PIM in order to create a gazetteer for entity lookup. The results of a similarity measure (1) involving the Levenshtein distance:

$$s_e = \frac{Levenshtein\,(e_i) - l_e}{l_e} \qquad (1)$$

—where s_e is the similarity score for an entity e_i, and l_e is its length—are sorted such that only those with the highest similarity are considered. Resource pairs with a score of above 0.9 are automatically interlinked, whereas pairs between 0.7 and 0.9 will require confirmation through the user interface, or discarded if ignored.

3.2.7 Hyperlink resolution

We make use of regular expressions to determine the type of the resource underneath hyperlinks embedded in microposts, i.e. image, video, audio, document, etc. Their type can also be determined by their content-type (e.g. *image/jpeg* generates an instance of an *dlpo:ImagePost*. In case the content-type is *text/html*, all boilerplate and template/presentation markup around the main textual content is detected and removed using *boilerpipe*. Depending on the type, the resource is then passed on to one of the pipelines for content analysis. The same keyword, topic and NE extraction techniques will here be applied to extract and annotate the relevant post subtype (e.g. *dlpo:ImagePost*, *dlpo:WebDocumentPost*) with topics and tags, as well as link them to matched resources in the PIM or the LOD cloud.

3.2.8 Time window analysis

A user's personal events are crawled from calendaring tools and services and integrated within the PIM as instances of *ncal:Event*. Events are an important source of information, since they usually occur at a specific time and have a limited duration. This conforms to what we refer to as a 'time window', which provides the analysis tasks with context information. In this sense, we have identified two tasks to which this extra context is especially beneficial: NER and co-reference resolution.

In microposts such as *"In the MSM workshop in Lyon. David is giving a great presentation, amazing work!"*, 'David' may be recognised as a named-entity (Person). To determine which specific *pimo:Person* instance this refers to without having more information is hard, if not impossible. Event instances in the PIM may refer to a list of attendees[20]. By taking into consideration a workshop event (instance of *ncal:Event*) that is known to be taking place now, and in which a particular David is also an attendee, will help disambiguate the micropost named-entity for David.

Microposts are usually informal and incomplete, e.g. *"great performance! I love them :)"*. Although no named-entities can be extracted from this text, by adding contextual information and performing co-reference resolution, the results may improve significantly. For example, the author of the above post is known to be attending an event described as "Chemical Brothers concert", where 'Chemical Brothers' has been recognised as a named-entity and disambiguated to the LOD resource `http://dbpedia.org/resource/The_Chemical_Brothers`. Time windowing analysis allows for co-reference resolution to be applied on both the post and the event description, such that 'them' in the former may be resolved to the entity in the latter, resulting in an additional semantic link between two items in the PIM.

4. USE CASES FOR APPLICATION

[20]As extracted from structured or unstructured calendar data—which is not discussed in this paper.

In this section we demonstrate the novelty of our approach through a simple use-case. Fig. 2 depicts how two items (larger rectangles), posted online by user Juan on two different accounts—Facebook and Twitter, are represented as DLPO instances, and integrated within the rest of the PIM ontologies. Once the post from Twitter (left-hand side) is obtained and transformed to RDF, core metadata is immediately generated for the *dlpo:LivePost* instance, consisting of relationships to the source (i.e. Twitter, as a known instance of *dao:Account*/*sioc:Container* in the PIM), to the creator (i.e. Juan Martinez, a known *pimo:Person*) and of other easily-obtained data, such as a timestamp. Since the post contains textual content that does not form part of a URI, an instance of *dlpo:Status* is generated as a subpost, through *dlpo:isComposedOf*. This subpost type only points to a string containing the textual content. Following a successful execution of the next stage in the semantic lifting process, the reference to "@aalford" is matched to the known PIM item 'Anna Alford' (*pimo:Person*). Similarly, the reference to '#Lyon' is matched to another existing item. This time however, the item is not found in the user's PIM, but in an external open KB (e.g. DBPedia). Since not enough information about the semantics of the relationship between the post and these items can be extracted (without as yet resorting to natural language processing techniques to determine e.g. that an arrival to a place amounts to a check-in), the items are loosely linked to the superpost through *dlpo:relatedResource*. Once the hyperlink from the post to the external Lyon representation is established, the same resource is adopted to the PIM as an external instance of *pimo:City*, thus virtually also becoming part of the PIM. In addition, the lemma for 'arrived', together with the reference to a *pimo:City*, are matched to two of the items (keyword "arrive", any location instance) assigned to a pre-defined *pimo:Topic* - Travel. As a result, this post is assigned this topic through *nao:hasTopic*.

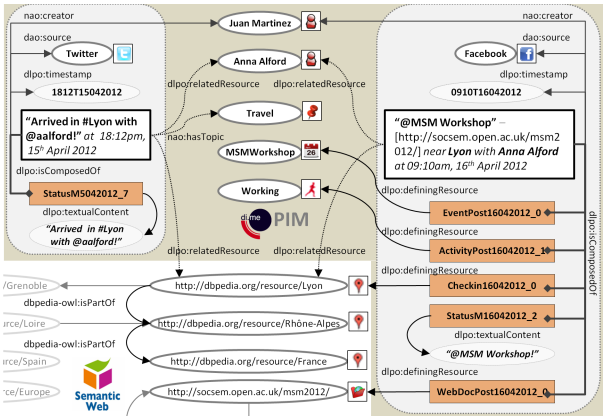

Figure 2: Semantic Integration of Post Concepts

The post retrieved from Facebook (right-hand side) generates another *dlpo:LivePost* instance. The relationship between the post and the 'Anna' and 'Lyon' PIM items is in this case easier to extract, since the information on this source is more structured (Anna is tagged, and Lyon is shown as a nearby location). The improved structure also translates into the generation of a specific *dlpo:Checkin* subpost instance for the superpost, whose defining resource is the same representation for Lyon linked to both external posts through *dlpo:relatedResource*. As the post on Facebook also includes a link, hyperlink resolution is employed to determine that it is not of any specific type (multimedia, blogpost, URI describing a resource, etc.). Thus, an instance of *dlpo:WebDocumentPost* is generated as a subpost, with its defining resource being the URL for the web document itself. It is also determined that the superpost is composed of two other subposts - an *dlpo:EventPost* and an *dlpo:ActivityPost*. The former is discovered after the 'MSM Workshop' named entity is matched to the label of an existing *pimo:Event* instance, thus automatically becoming the subpost's defining resource. Similarly, lemmas from the textual content of the superpost are matched to keywords attached to an existing system-defined instance of the *dpo:Activity* class in the DPO—'Working'. As a result, it is established that the user is posting about a 'Working' activity that is currently in progress. As shown in Fig. 2, our approach enables posts from different sources to be mapped to unique representations of items in both the PIM and the LOD cloud.

5. FUTURE WORK AND CONCLUSIONS

In this paper we have presented an approach for analysing and extracting presence-related information from Social Web sharing activities, in order to enrich a user's integrated Personal Information Model, so as to improve the user's experience in using a proposed intelligent personal information management system—the di.me userware. Apart from being the basis for providing context-aware recommendations, aggregated user presence-related information also becomes more readily available for Social Web sharing.

Our main contribution is the DLPO ontology, which successfully combines aspects from both user presence and online posting domains in a concise ontology, itself integrated within established PIM Knowledge Models. Although most of the discussed techniques for semantic lifting have already been implemented, some in a more advanced state (keyword/topic extraction, hyperlink resolution) than others (NEE, NER & co-reference resolution), the overall approach is still being improved and extended. Apart from discussed advanced features such as time window analysis, next in line for investigation are techniques for a richer semantic analysis of posts, either by additional natural language processing techniques (e.g. to discover implicit user actions in text) or through the exploitation of metadata that is already attached to shared items (e.g. location coordinates from a posted image), through the use of vocabularies such as Rich Snippets, Open Graph protocol, schema.org, RDFa, Microdata and Microformats. This metadata is much more easily-obtained, since it already contains explicit references to entities in the LOD Cloud. Once all of the envisaged functionality is in place and the di.me userware is deployed for use, we will perform an adequate evaluation of our entire approach as we propose it. Evaluation will be performed on three different aspects — i) success of online post decomposition compared against a manual approach, ii) determination of PIM concepts that online posts are mostly/rarely/never linked to, and iii) decomposition of online posts within different social networks, to find out which subtypes are mostly useful for users e.g. checkins - users can see the current location of their contacts in a graph format, event posts - are

automatically stored in a personal calendar, activity posts - graph representation showing the activities that a user normally does. In the meantime, we plan to evaluate individual aspects of the overall technique and improve it accordingly.

6. ACKNOWLEDGMENTS

This work is supported in part by the European Commission under the Seventh Framework Program FP7/2007-2013 (*digital.me* – ICT-257787) and in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (*Líon-2*).

7. REFERENCES

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing User Modeling on Twitter for Personalized News Recommendations. In *Proceedings of the International Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 1–12, 2011.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[3] J. G. Breslin, A. Harth, U. Bojars, and S. Decker. Towards semantically-interlinked online communities. In *European Semantic Web Conference (ESWC)*, Lecture Notes on Computer Science, pages 500–514. Springer, 2005.

[4] D. Brickley and L. Miller. Foaf vocabulary specification 0.98, 2010.

[5] A. E. Cano, A.-S. Dadzie, V. S. Uren, and F. Ciravegna. Sensing presence (presense) ontology: User modelling in the semantic sensor web. In *Proceedings of the ESWC Workshop on User Profile Data on the Social Semantic Web (UWeb 2011)*, 2011.

[6] A. E. Cano, S. Tucker, and F. Ciravegna. Capturing entity-based semantics emerging from personal awareness streams. In *Proceedings of the 1st Workshop on Making Sense of Microposts (#MSM2011) at ESWC*, pages 33–44, 2011.

[7] I. Celino, D. Dell'Aglio, E. D. Valle, Y. Huang, T. Lee, S. Park, and V. Tresp. Making sense of location based micro-posts using stream reasoning. In *Proceedings of the 1st Workshop on Making Sense of Microposts (#MSM2011) at ESWC*, pages 13–18, 2011.

[8] J. Chang and E. Sun. Location: How users share and respond to location-based data on social networking sites. In *Proceedings of the AAAI Conference on Weblogs and Social Media*, pages 74–80, 2011.

[9] S. Choudhury and J. Breslin. Extracting semantic entities and events from sports tweets. In *Proceedings of the 1st Workshop on Making Sense of Microposts (#MSM2011) at ESWC*, pages 22–32, 2011.

[10] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[11] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. *Text Processing with GATE (V.6)*. 2011.

[12] S. Deerwester. Improving Information Retrieval with Latent Semantic Indexing. In *Proceedings of the 51st ASIS Annual Meeting (ASIS '88)*, Atlanta, Georgia, 1988. American Society for Information Science.

[13] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Discov.*, 8(1):53–87, Jan. 2004.

[14] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 216–223, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[15] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 359–367, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[16] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004.

[17] A. Miles and S. Bechhofer. Skos simple knowledge organization system reference, 2009.

[18] A. Passant, U. Bojars, J. Breslin, T. Hastrup, M. Stankovic, and P. Laublet. An overview of smob 2: Open, semantic and distributed microblogging. pages 303–306, 2010.

[19] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the 13th Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

[20] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named Entity Recognition in Tweets: An Experimental Study. In *EMNLP*, 2011.

[21] O. Sacco and A. Passant. A privacy preference ontology (ppo) for linked data. In *Proceedings of the Linked Data on the Web Workshop (LDOW)*, 2011.

[22] M. Sintek, S. Handschuh, S. Scerri, and L. van Elst. Technologies for the social semantic desktop. In *Reasoning Web. Semantic Technologies for Information Systems*, Lecture Notes in Computer Science, pages 222–254. Springer-Verlag, 2009.

[23] M. Stankovic and J. Jovanovic. Online presence in social networks. In *Proceedings of the W3C Workshop on the Future of Social Networking*, 2009.

[24] T. Steiner, A. Brousseau, and R. Troncy. A tweet consumers' look at twitter trends. In *Proceedings of the 1st Workshop on Making Sense of Microposts (#MSM2011) at ESWC*, 2011.

[25] W. Wu, B. Zhang, and M. Ostendorf. Automatic generation of personalized annotation tags for twitter users. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 689–692, Stroudsburg, PA, USA, 2010.

[26] K. Zoltan and S. Johann. Semantic analysis of microposts for efficient people to people interactions. In *Proceedings of the 10th Roedunet International Conference (RoEduNet), 2011*, pages 1–4, 2011.