Discovering Semantic Equivalence of People behind Online Profiles

Keith Cortis, Simon Scerri, Ismael Rivera, and Siegfried Handschuh

Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland
firstname.surname@deri.org

Abstract. Users are currently required to create and separately manage duplicated personal data in heterogeneous online accounts. Our approach targets the crawling, retrieval and integration of this data, based on a comprehensive ontology framework which serves as a standard format. The motivation for this integration is to enable single point management of the user's personal information. The main challenge faced by this approach is the discovery of semantic equivalence between contacts described in online profiles, their attributes and shared posts. Contacts found to be semantically equivalent to persons that are already represented within the user's personal information model are linked together. In this paper we outline our part-syntactic, part-semantic approach to online profile integration, the current status and future plans for research and development concerned with this challenge.

Keywords: semantic equivalence, online profile, personal information model, ontologies, social networks, semantic lifting, semantic web

1 Introduction

At present, the typical computer literate user is forced to create a personal profile for each online account they would like to use. Since the recent shift towards the usage of remote data management and sharing services, this necessity has become even more pressing. Popular online accounts now vary from general social networking platforms to specific email, instant messaging, calendaring, task management and file-sharing services as well as business-oriented customer management services. Personal data in these accounts ranges from the more static identity-related information, to more dynamic information about one's social network as well as physical and online presence. In the context of this paper, we refer to all these kinds of personal data, stored on one of many distinguishable online accounts, as a user's 'online profile'.

The current situation results in personal data being unnecessarily duplicated over different platforms, without the possibility to merge or port any part of it [2], thus forcing users to also manage this data separately and manually. This is reflected in a survey[1] that we conducted, where 16% *always*, 20% *frequently* and 38% *sometimes* use the same personal information within their 'business' (e.g. professional networks) and 'administrative' (e.g. e-commerce) profiles. On the other hand, the 'social/private'

[1] http://smile.deri.ie/node/517

profiles of 12% *always*, 6% *frequently* and 40% *sometimes* contain the same personal information as their business/administrative profiles. Our aim is to enable the user to create, aggregate and merge multiple online profiles into one digital identity, through the di.me[2] userware - a single access point to the user's personal information sphere [25]. The latter also refers to personal data on a user's multiple devices (e.g. laptops, tablets, smartphones). This makes the di.me userware sophisticated and novel since it does not only 'attack' the distributed/duplicated online profile management problem, but targets the integration of distributed/duplicated personal information found across multiple local and remote sources. The already integrated data is stored in the user's personal information sphere which holds all the valuable information of the user on a personal server. The advantage of creating a digital identity within the di.me userware is that of automatically integrating several identities into one with no, or minimal, user effort. This would then enable easier management of the multiple identities, without expecting existing systems to adopt our model or the user to do the integration manually, both of which aren't practical. Results from our survey also outline that 32.7% would *extremely* favour a system that synchronises and shows you personal information collected from different personal online sites, 26.5% favour the idea *quite a bit*, whilst 20.4% are *moderately* in favour. Additionally, 30.6% would *extremely* favour a system that enables you to centrally modify and update your information in different personal sites from one location, 30.6% favour the idea *quite a bit*, whilst 14.3% are *moderately* in favour. These statistics motivate the development of the di.me userware.

In this paper we will only focus on the integration of heterogeneous online user profiles, a task which is not straightforward for two main reasons. First, no common standards exists for modelling profile data in online accounts [20], making the retrieval and integration of federated heterogeneous personal data instantly a hard task. A second problem is that the nature of some of the personal data on digital community ecosystems [13], such as known contacts (resources) and presence information, is dynamic. To address these difficulties, we propose the use of a standard format that is able to handle both the more static as well as dynamic profile data. This comes in the form of an integrated ontology framework consisting of a set of re-used, extended as well as new vocabularies provided by the OSCA Foundation (OSCAF)[3] (only the most relevant ontologies will be mentioned in this paper). Our approach is to map and integrate various online user profiles onto this one standard representation format. The first stage of this approach discovers semi/unstructured information by crawling attributes that are available through online account APIs, resulting in a separate representation for each respective online profile. These representations maintain links to the source account as well as to the external identifiers of the specific online profile attributes. Additionally, all crawled attributes, in our case the profile information, are aggregated into what we refer to as the user's 'super profile'. The second stage of our approach targets the mapping of attributes for each of the represented online profiles with equivalent attributes for the super profile. The use of ontologies and Resource Description Framework (RDF)[4] as the main data representation means that the mapping we pursue considers both syn-

[2] http://www.dime-project.eu/
[3] http://www.oscaf.org/
[4] http://www.w3.org/RDF/

tactic as well as semantic similarities in between online profile data. Our approach is performing semantic lifting and not traditional ontology matching since we are discovering resources from a user's profile (schema of particular online account) which are then mapped to our ontology framework. We then attempt to discover semantic equivalence between persons (this includes both the user and their contacts) that are known in multiple online accounts, based on the results of individual attribute matching. An appropriate semantic equivalence metric is one of the requirements for aspiring self-integrating system [21], such as the di.me userware.

Several techniques may be required in-order to discover if two or more online persons are semantically equivalent. The most popular techniques are syntactic based, i.e. a string/value comparison is performed on the various person profile attributes. Our ontology-based approach allows us to extend the matching capabilities 'semantically', ensuring more accurate results based on clearly-specified meanings of profile attribute types, as well as through an exploration of their semantic (in addition to syntactic) relatedness. The discovery of semantically equivalent person representations results in their semantic integration at the Personal Information Model (PIM) level of the user's data. The PIM handles unique personal data that is of interest to the user such as the user's singular digital identity, files, task lists and emails. It is an abstraction of the possibly multiple occurrences of the same data as available on multiple online accounts and devices. The users have complete control over their accessed personal data, since our approach does not target sharing of personal information. In the remainder of the paper we start by discussing and comparing related work in Section 2. Details on our approach are then provided in Section 3. An update of the current status and prototype implementation is then provided in Section 4, before a list of our targeted future aspirations and a few concluding remarks in Section 5.

2 Related Work

The process of *matching* takes two schemas or ontologies (each of which is made up of a set of discrete entities such as tables, XML[5] elements, classes, properties, etc.) as an input, producing relationships (such as equivalence) found between these entities as output [26]. COMA++ [3] is one of the most relevant schema and ontology matching tools that finds out the semantic correspondences among meta-data structures or models. Given that these matching problems are overcome, it would benefit service interoperability and data integration in multiple application domains. Several techniques and prototypes were implemented in-order to solve the matching problem in a semi-automatic manner such as Falcon-AO (ontology matching) [14], thus reducing manual intervention. Our approach is different to the mentioned traditional approach since we aren't concerned with matching two conceptualisations (schemas or ontologies), but a schema of an online account e.g. a social network to an ontology or set of ontologies. We refer to this process as semantic lifting, since we are lifting semi/unstructured information (the user's profile attributes) from a schema as discussed in Section 3.1, which is manually mapped to an interoperable standard (ontology framework) as discussed in Section 3.2.

[5] http://www.w3.org/XML/

Findings in [15] suggest that provided enough information is available, multiple user profiles can be linked at a relatively low cost. In fact, their technique produces very good results by considering a user's friends list on multiple online accounts. Earlier approaches rely on just a specific Inverse Functional Property (IFP) value e.g. email address or name [17],[12]. However, as pointed out in [5], IFP assumptions are rather shallow, since users are able to create multiple accounts even within the same social network (e.g. a business-related profile, social profile, etc.) each time using different identification, e.g. email addresses.

A number of approaches rely on formal semantic definitions, through the use of ontologies and RDF, to enable portability of online profiles. The work by [22] presents an online application that transforms a user's identity on a social network (Facebook) into a portable format, based on the Friend of a Friend (FOAF) ontology [6]. The approach described in [20] goes on step forward, attempting to integrate multiple online profiles that have been converted to FOAF. As opposed to IFP approaches, this approach takes into consideration all (FOAF) profile attributes, assigning different importance levels and similarity measures to each. Although FOAF enables a much richer means for profile attribute comparison, we use a more comprehensive conceptualisation through the Nepomuk Contact Ontology (NCO) [19], which is integrated into a comprehensive ontology framework. This integration enables attributes in multiple profiles to be semantically related to unique, abstract representations in the user's PIM. Once the technique in [20] sees the profiles transformed to a FOAF representation, a number of techniques are used for syntactic matching between short strings and entire sentences. In addition, the syntactic-based aspect of our matching will also perform a Linguistic Analysis to yield further information about the typed profile attributes. Named Entity Recognition (NER) can discover more specific types than the ones known (e.g. identifying city and country in a postal address) and recognise abbreviations or acronyms in attribute labels.

Many approaches enhance the otherwise syntactic-based profile matching techniques with a semantic-based extension. In particular, the above-cited work by Raad et. al. is supplemented with an Explicit Semantic Analysis [11], whose aim is to detect semantic similarity between profile attributes through the computation of semantic relatedness between text, based on Wikipedia. A similar approach [27] uses snippets returned from an online encyclopedia to measure the semantic similarity between words through five web-based metrics. Our approach will consider semantic relatedness to determine similarity between entities not only based on their labels or values, but also on a semantic distance to other relevant concepts. For example, although an address in one profile might consist of just the city, and another address might refer to only the country, the fact that the city in the first profile is known to be in the country defined for the second profile will be considered as a partial match.

The calculation of such measures within different systems or domains is a very important task, given the increase in access to heterogeneous and independent data repositories [9]. Research efforts conducted by [28] identify three common approaches for calculating the semantic distance between two concepts, namely i) the knowledge-based approach which uses remote Knowledge Bases (KBs) such as WordNet[6] (count edge distance) [7], ii) lexico-syntactic patterns (makes binary decisions), and iii) statisti-

[6] http://wordnet.princeton.edu/

cal measures (uses contextual distributions or concept co-occurrences). The mentioned techniques are not relevant for certain cases, as the concept distances cannot be calculated. This means that such a process is not straightforward, especially if a personal KB is used, where a good distance metric needs personal adjustments in-order to cater for a particular user's needs. Normally for a personal ontology (can be domain specific), in our case the PIM, several concepts are not available within remote KBs. Therefore, it's impossible to calculate the semantic distance between two concepts, if remote KBs are used alone. Hierarchical semantic KBs such as the ones constructed on an "is a" structure, can be fundamental for finding the semantic distance between concepts [16].

There is one major distinction between our approach and the semantic-based approaches described above. Although remote KBs such as DBpedia[7] are to be considered as a backup, the KB on which we initially perform a similarity measure is the user's own PIM. The PIM is populated partly automatically - by crawling data on the user's devices, applications and online accounts, and partly by enabling the user to manually extend the representations of their own mental models. The advantage here is that the PIM contains information items that are of direct interest to the user, and is thus more relevant to the user than external structured or partly structured KBs. Therefore, the semantic matching of profiles is bound to yield more accurate results, based on a KB that is more personal and significantly smaller.

3 Approach

Our online profile (instance) matching approach will involve four successive processes (A-D), as outlined by Fig. 1 and discussed below.

3.1 Retrieval of User Profile Data from Online Accounts

The first step is to retrieve personal information from various online accounts such as Facebook, Twitter and LinkedIn, and is fairly straightforward once the required API calls are known. We target several categories of online profile data such as the user's own identity-related information, their online posts, as well as information about the user's social network, including the identities and posts shared by their contacts.

3.2 Mapping User Profile Data to the Ontology Framework

Once online profile data has been retrieved from an online account, it is mapped to two particular ontologies in our ontology Framework. Identity-related online profile information is stored as an instance of the NCO Ontology, which represents information that is related to a particular contact. The term 'Contact' is quite broad, since it reflects every bit of data that identifies an entity or provides a way to communicate with it. In this context, the contact can also refer to the user's own contact information. Therefore, both the user and their contacts as defined in an online profile are represented as instances of *nco:Contact*. Presence and online post data for the user is stored as instances of the LivePost Ontology (DLPO)[8], a new ontology for the representation of

[7] http://dbpedia.org/

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

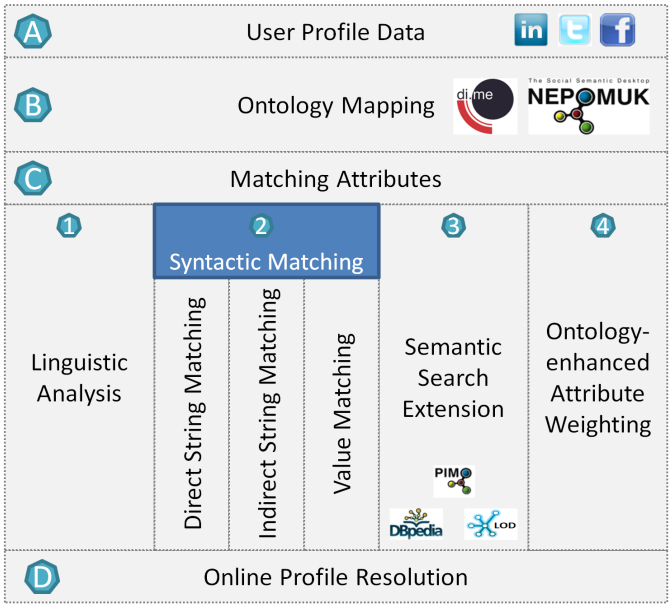

Fig. 1. Approach Process

dynamic personal information that is popularly shared in online accounts, such as multimedia posts (video/audio/image), presence posts (availability/activity/event/checkin), messages (status messages/comments) and web document posts (note/blog posts).

Fig. 2 demonstrates how the above ontologies can be used to store online profile data from an online account (OnlineAccountX). The figure also shows the user's super profile (di.meAccount). An explanation of how the other ontologies in the framework can be used to effectively integrate the two profiles once semantic equivalence is discovered, is provided later on. The upper part of the figure refers to the T-box, i.e. the ontological classes and attributes, whereas the lower part represents the A-box, containing examples of how the ontologies can be used in practice (straight lines between the A- and T-box denote an instance-of relationship).

The attributes of the online user profiles will be mapped to their corresponding properties within our ontology framework. The example shows five identity-related profile attributes that have been mapped to the NCO (affiliation, person name, organisation, phone number, postal address). Presence-related profile information is also available in the form of a complex-type 'livepost', consisting of a concurrent status message - "Having a beer with Anna @ESWC12 in Iraklion", a check-in (referring to the *pimo:City* representation for Heraklion through *dlpo:definingResource*) and an event post (referring to the *pimo:Event* instance representing the conference through *dlpo:definingResource*). *dlpo:definingResource* defines a direct relationship between a 'livepost' subtype and a PIM item. A person, "Anna" is also tagged in this post, as referred by *dlpo:relatedResource*. This property creates a semantic link between a 'livepost' and the relevant PIM items.

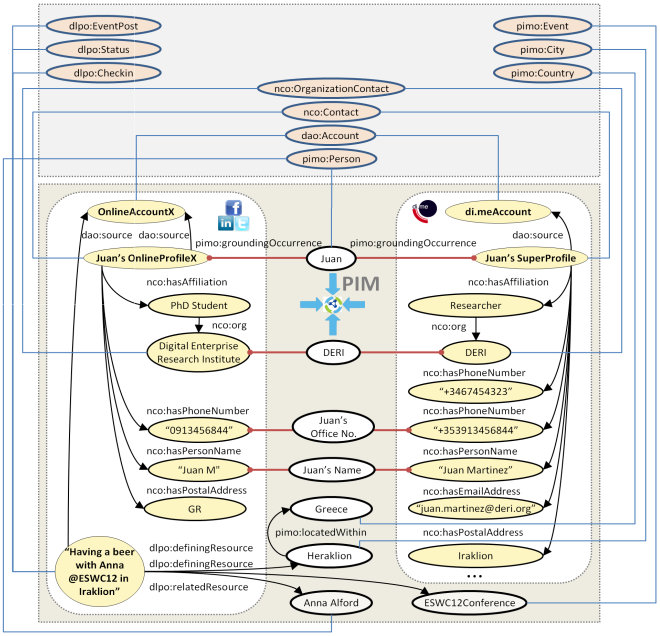

Fig. 2. Approach Scenario

3.3 Matching User Profile Attributes

Our approach towards matching the user profile attributes (metadata matching), considers the data both at a semantic and syntactic level. It involves four successive processes as outlined within the third level (C) of Fig. 1.

Linguistic Analysis Once the transformation and mapping of the user's profile data to its RDF representation has been performed, a matching process is initiated against the user's PIM in order to find similar attributes or links and relations between them. In the case that the profile attribute is known to contain an atomic value (e.g. a person's name, phone number, etc.), no further linguistic analysis is performed. However, profiles attributes may contain more complex and unstructured information such as a postal address (e.g. "42 Upper Newcastle Road, Lower Dangan, Galway, Ireland"). For such attributes, a deeper linguistic analysis is required to discover further knowledge from their values; concretely, a decomposition into different entities or concepts is the goal pursued. In the postal address example, the aim is to find out that '42 Upper Newcastle Road' refers to the most specific part of the address information, 'Lower Dangan' to an area or district, 'Galway' to a city and 'Ireland' to a country. The techniques applied to extract or decompose the attribute values are regular expressions and gazetteer lookups. Typically both techniques work well when the domain or structure is known. Therefore, the algorithm distinguishes profile attributes by type or nature, which is known at

this stage, to apply different regular expressions and use different gazetteers. Abbreviations and acronyms are also covered in this analysis by including entries for them in the gazetteers (e.g. a gazetteer for countries also includes the ISO 3166 codes).

Finally, there are special profile attributes which let the user describe themselves, or even include hyperlinks to their personal websites. The text in these attributes is also analysed by a Natural Language Processing pipeline in order to extract named-entities and perform the proper lookups in the gazetteers. However, users normally just provide a hyperlink to their personal website[9], due to size limitations for the description attribute, or the reluctance of the users to enter such information within all their online profiles. In such cases, the hyperlink is resolved and its content is extracted for further analysis, mainly to discover any named-entities (e.g. city) that will enrich the description attribute. Despite of not having these meaningful links between the entities and the profile, such information is used in order to re-balance the weights of certain attributes as described in the Ontology-enhanced Attribute Weighting sub-section below. For example, a Twitter user who only provides a username in her profile would not be able to match to a richer profile which contains her name, postal address, phone number, email, etc., if the username is unknown. However, if this information is found in the form of entities within her personal website, the likelihood that two profiles match increases considerably.

Syntactic Matching Straightforward **value matching** is applied on attributes that have a non-string literal type (e.g. birth date or geographical position), since these have a strict, predefined structure. For attributes of type string (*xsd:string*), if their ontology type (e.g. person name, postal address) is either known beforehand or discovered through NER, **direct string matching** is applied. In both cases, the matching takes as input the attribute in consideration against PIM instances of a similar type. For example, in Fig. 2., the label of the organisation (*nco:OrganizationContact* instance) specified within the *nco:org* property for the user's online account profile (i.e. 'Digital Enterprise Research Institute') is matched against other organisation instances within the PIM. The super profile instance 'DERI' is one example of other PIM instances having the same type. The fact that in this case one of these two equivalent profile organisation attributes is an acronym for the other one will be taken into consideration by the employed string matching technique. In the event that the attribute entity remains unknown even after NER is performed, **indirect string matching** is applied over all PIM instances, regardless of their type.

A string matching metric is used for syntactically matching user profile attribute values that are obtained from an online account to attribute values that are stored in the PIM KB. The recursive field matching algorithm proposed by Monge and Elkan [18] is applied for matching string values. A degree of 1.0 signifies that string 'A' and string 'B' fully match or one string abbreviates the other. On the other hand, a degree of 0.0 signifies that there is no match between two strings. All sub-fields of string 'A' and string 'B' are also compared against each other. The Monge-Elkan string matching metric (1), considered as one of the most accurate [8], is defined as follows:

[9] In an analysis of the 'description' field for 125 Twitter users, we found that 54% were linking to a web page that contains personal information about them.

$$match\,(A,B) = \frac{1}{|A|}\sum_{i=1}^{|A|}\max\{sim'\,(A_i,B_j)\}_{j=1}^{|B|} \quad (1)$$

—where *sim'* is a particular secondary distance function that is able to calculate the similarity between two tokens 'A' and 'B'. The major reason for choosing this metric is that it holds for matching an attribute value to its abbreviation or acronym, unlike other metrics considered such as Levenshtein distance, Jaro and Jaro-Winkler. This technique considers the abbreviation that is either: i) a prefix, ii) a combination of a prefix and suffix, or iii) an acronym, of the expanded string, or else a concatenation of prefixes from the expanded string. Our plan is to extend this metric to match non-trivial variations of an expanded string e.g. username 'ramauj' to the full name 'Juan Martinez'.

Semantic Search Extension Once the syntactic matching is complete, a semantic search extension process follows. Referring again to our example, the user's address known for the super profile (di.meAccount) is listed as 'Iraklion', and is related to an instance of a *pimo:City*, 'Heraklion'. The one just retrieved from the online account profile (OnlineAccountX) refers to 'GR', which is found to be related to a particular instance of *pimo:Country*, 'Greece'. Although the two address attributes do not match syntactically, they are semantically related. Given that the profile in question is the user's, it is highly likely that through some other data which is either automatically crawled or enriched by the user, the PIM contains references to both these locations, and that semantic relationships exist in between. In the example, through the PIM KB, the system already knows that the city and country instances related to both addresses are in fact related through *pimo:locatedWithin*. This constitutes a partial semantic match, to be taken into consideration when assigning semantic-based attribute weights. If such data didn't exist within the PIM KB (main KB for matching), remote KBs such as DPBedia or any other dataset that is part of the Linked Open Data cloud[10], will be accessed to determine any possible semantic relationship. Another example centres around Juan's two roles, listed as a 'Researcher' for 'DERI' within his super profile, and as a 'PhD Student' for 'Digital Enterprise Research Institute' on his online account. Although less straightforward, a semantic search here would largely support the syntactic search in determining that there is a high match between these two profile attributes, after finding that DERI is a research institute which employs several Researchers and PhD students.

Ontology-enhanced Attribute Weighting To discover semantic equivalence between persons in online profiles or otherwise, an appropriate metric is required for weighting the attributes which were syntactically and/or semantically matched. Factors that will be taken into account by the metric are the total number of attributes that were mapped to our ontology framework, the number of syntactically matched attributes, the number of attributes matched based on the semantic search extension, and the importance of attributes depending on the target domain of the specific online account. In addition, ontology-enhanced attribute weights are an added benefit of our ontology framework

[10] http://lod-cloud.net/

over other ontologies such as FOAF. Attribute constraints defined in the NCO ontology, such as cardinality and inverse functional properties, enable the assignment of different predefined weights to the attributes. Thus, the properties that have a maximum or an exact cardinality of 1 have a higher impact on the likelihood that two particular profiles are semantically equivalent. Carrying even a higher predefined weight are inverse functional properties, which uniquely identify one user. Examples of attributes having cardinality constraints are first name, last name and date of birth, whereas an example of a inverse functional property is a private email address or a cell phone number. Profile attributes such as affiliation, organisation, city and country have no such cardinality constraints defined in the ontology, and as a result they have a lower weight.

3.4 Online Profile Matching

Based on the attribute weighting metric, we define a threshold for discovering semantic equivalence between elements of a user's online profiles, i.e. personal identity and information that is already known and represented at the PIM level. A user can then be suggested to merge all kinds of profile information, e.g. their 'organisation' from various online profiles into their super profile, depending on this threshold. This includes marking contacts for the same unique person as 'known' over multiple online accounts.

The actual integration of semantically-equivalent personal information across distributed locations is realised through the 'lifting' of duplicated data representations onto a more abstract but unique representation in the PIM. The Personal Information Model Ontology (PIMO) [23] provides a framework for representing a user's entire PIM, modelling data that is of direct interest to the user. By definition, PIMO representations are independent of the way the user accesses the data, as well as their source, format, and author. Initially, the PIM will be populated with any personal information that is crawled from a user's particular online account or device. Therefore, if there is no match of a particular entity, a new instance is created. In the example shown in Fig. 2, Juan's PIM (grey area) 'glues' together all the things he works with uniquely, irrespective of their multiple 'occurrences' on different devices and/or online accounts. First and foremost, the PIM includes a representation for the user himself, as a *pimo:Person* instance. This instance refers to the two shown profiles through the *pimo:groundingOccurrence* property, which relates an 'abstract' but unique subject to one or more of its occurrences. For example, the unique *pimo:City* instance has multiple occurrences in multiple accounts, and is related to both Juan's postal address and his check-in as defined on his online account. The advantage of using ontologies is evident here - resources can be linked at the semantic level, rather than the syntactic or format level. For example, although the user's name or organisation differ syntactically, the discovery that they are semantically equivalent is registered within the PIM.

4 Implementation

This section describes the development progress so far. The current prototype employs the Scribe OAuth Java library[11] to retrieve data from a LinkedIn profile. Scribe supports

[11] https://github.com/fernandezpablo85/scribe-java

Table 1. Ontology Mapping of LinkedIn Attributes

Query	Attribute	Ontology Mapping
http://api.linkedin.com/v1/people/∼:	id	$\xrightarrow{nao:externalIdentifier}$ <value>
(id,first-name,last-name,location:	first-name	$\xrightarrow{nco:hasPersonName}$ nco:PersonName $\xrightarrow{nco:nameGiven}$ <value>
(name),picture-url,summary,	last-name	$\xrightarrow{nco:hasPersonName}$ nco:PersonName $\xrightarrow{nco:nameFamily}$ <value>
positions,phone-numbers,	location:(name)	$\xrightarrow{nco:hasLocation}$ geo:Point $\xrightarrow{nao:prefLabel}$ <value>
im-accounts,date-of-birth)	picture-url	$\xrightarrow{nco:photo}$ <value>
http://api.linkedin.com/v1/people/∼	summary	$\xrightarrow{nao:description}$ <value>
/connections:(id,first-name,last-	positions	$\xrightarrow{nco:hasAffiliation}$ nco:Affiliation $\xrightarrow{nco:title/role/department/org}$ <value>
name,location:(name),picture-url,	phone-numbers	$\xrightarrow{nco:hasPhoneNumber}$ nco:PhoneNumber $\xrightarrow{nco:phonenumber}$ <value>
summary,positions,phone-numbers,	im-accounts	$\xrightarrow{nco:hasIMAccount}$ nco:IMAccount $\xrightarrow{nco:imAccountType/imID}$ <value>
im-accounts,date-of-birth)	date-of-birth	$\xrightarrow{nco:hasBirthDate}$ nco:BirthDate $\xrightarrow{nco:birthdate}$ <value>
http://api.linkedin.com/v1/people/∼:	id	$dlpo:LivePost \xrightarrow{nao:externalIdentifier}$ <value>
(current-share)	timestamp	$dlpo:LivePost \xrightarrow{dlpo:timestamp}$ <value>
	comment	$dlpo:LivePost \xrightarrow{dlpo:textualContent}$ <value>
	source name	$dlpo:LivePost \xrightarrow{dao:source}$ dao:Account $\xrightarrow{nao:prefLabel}$ <value>

major 1.0a and 2.0 OAuth APIs such as Google, Facebook, LinkedIn, Foursquare and Twitter, and can thus be used to extend future profiles.

Table 1 shows different types of LinkedIn service calls that our prototype supports (column one). The first retrieves a user's profile data, whereas the second retrieves a user's contact profile data. The third query retrieves status updates from the user. The calls return a set of LinkedIn profile data for the user or their connections, of which we currently map the shown list (column two) to the specific concepts and properties in our ontology framework (column three). The first set of ontology properties in the third column are attached to the *nco:Contact* instance representing the user or one of their contacts (omitted from the Table), whereas the second set of ontology properties are attached to the respect *dlpo:LivePost* instance. Both instances are linked to the online account from which they where retrieved via *dao:source*, this case being a representation of the LinkedIn online account.

Since the LinkedIn API[12] data is returned in XML, we required a transformation of this data into an RDF representation, for mapping to our ontologies. The translation between XML to RDF is quite a tedious and error-prone task, despite the available tools and languages. Although an existing approach is to rely on Extensible Stylesheet Language Transformations (XSLT)[13], the latter was designed to handle XML data, which in contrast to RDF possesses a simple and known hierarchical structure. Therefore, we use the XSPARQL [1] query language. XSPARQL (W3C member submission) provides for a more natural approach based on merging XQuery[14] and SPARQL[15] (both W3C Recommendations). The transformation between the XML LinkedIn data into our RDF representation (using Turtle[16] as the serialization format) is declaratively expressed in a

[12] https://developer.linkedin.com/rest
[13] http://www.w3.org/TR/xslt
[14] http://www.w3.org/TR/xquery/
[15] http://www.w3.org/TR/sparql11-query/
[16] http://www.w3.org/TR/turtle/

XSPARQL query, which also covers the transformation of profile data from any social network adherent to the OpenSocial standard[17].

For the linguistic analysis and NER process, presented in Section 3.3, we have selected the General Architecture for Text Engineering (GATE)[18] platform, which allows decomposing complex processes—or *'pipeline'*—into several smaller tasks or modules with a specific purpose or using a specific (even third-party) algorithm. GATE is is distributed with the ANNIE information extraction system [10], which includes a variety of algorithms for sentence analysis and pre-defined gazetteers for common entity types (e.g. countries, organizations, etc.), which we extended with acronyms or abbreviations where necessary. We employ the GATE *Large KB Gazetteer* module in order to make use of the information stored within the user's PIM, since it can get populated dynamically from RDF data.

Listing 1.1 shows an example of online profile data retrieved from the LinkedIn account for user "Juan Martinez". The RDF representation (in Turtle syntax) shows how the data is mapped to our ontology framework, through the XSPARQL transformer. The LinkedIn account representation (_:acct1 as an instance of *dao:Account*) contains references to two contacts known within (_:c1, _:c2 as instances of *nco:Contact*), one of which (_:c1) is the Juan's own contact representation. Shown attached to Juan's contact instance is a series of identity-related information as well as one status message post (instance of *dlpo:Status*). This example highlights the comprehensiveness of our integrated ontology framework in dealing with various types of online profile data, when compared to other integrated ontology approaches such as the use of FOAF and Semantically-Interlinked Online Communities [4]. More importantly, it also illustrates how integration of online profile data is achieved at the semantic level. Once the two contacts in the online profile (including the one for the user) are discovered to be semantically equivalent to persons that are already represented in the PIM, a link is created between them through *pimo:groundingOccurrence*. The PIM Metadata at the bottom of Listing 1.1 demonstrates how the same unique person representations at the level of the PIM can point to multiple occurrences for that person, e.g. contacts for that person as discovered in online accounts, including the ones just retrieved from LinkedIn.

5 Conclusions and Future Work

In this paper, we discuss the possibility of eliminating the need for the user to separately manage multiple digital identities in unrelated online accounts. Our approach targets the crawling and retrieval of user profile data from these accounts, and their mapping onto our comprehensive ontology framework, which serves as a standard format for the representation of profile data originating from heterogeneous distributed sources. Our main target is the discovery of semantic equivalence between contacts described in online profiles, through a metric which computes a weighted semantic similarity of their individual attributes. Aggregated profile data is lifted onto a unique PIM representation and integrated in a super profile. The integrated data in the PIM is the main KB for matching since it's personalised and thus contains the most valuable information about

[17] http://code.google.com/apis/opensocial/
[18] http://gate.ac.uk/

```
#LinkedIn Profile Metadata
_:acct1 a dao:Account .
_:acct1 nao:prefLabel "LINKEDINAccount" .
_:c1 a nco:Contact .
_:c1 dao:source _:acct1 .
_:c1 nao:externalIdentifer "J7qb-67bTP" .
_:c1 nco:hasPersonName _:cn12 .
_:cn12 a nco:PersonName .
_:cn12 nco:nameGiven "Juan" .
_:cn12 nco:nameFamily "Martinez" .
_:c1 nco:hasAffiliation _:pos8 .
_:pos8 a nco:Affiliation .
_:pos8 nao:externalID 224093780 .
_:pos8 nco:role "Strategy Manager" .
_:pos8 nco:start "2003-1-1T00:00:00Z" .
_:pos8 nco:org _:org16 .
_:org16 a nco:OrganizationContact .
_:org16 nie:title "Ingeneria Ltd." .
...
_:stms644819790 a dlpo:Status .
_:stms644819790 dao:source _:acct1 .
_:stms644819790 nao:externalIdentifier "s6448190" .
_:stms644819790 dlpo:timestamp "2011-10-26T21:32:52" .
_:stms644819790 dlpo:textualContent "Seeking Job" .
...
_:c2 a nco:Contact .
_:c2 dao:source _:acct1 .
_:c2 nco:hasPersonName _:cn22 .
_:cn22 a nco:PersonName .
_:cn22 nco:nameGiven "Anna" .
_:cn22 nco:nameFamily "Alford" .
...

#PIM Metadata
_:PIM a pimo:PIM .
_:PIM pimo:creator _:user .
_:user a pimo:Person .
_:user pimo:groundingOccurrence _:c1 .
_:user pimo:groundingOccurrence _:c23 .
_:user pimo:groundingOccurrence _:c18 .
...
_:user foaf:knows _:person35 .
_:person35 pimo:groundingOccurrence _:c2 .
_:person35 pimo:groundingOccurrence _:c53 .
...
```

Listing 1.1. User Profile Transformer Output and PIM Integration

the user. One of the motivations for the di.me userware (currently in development), is to enable the user a single entry-point to distributed personal information management. This would enable easier management for the user, where no or minimal user effort would be required for the integration of such personal information.

The current prototype is able to retrieve a user's profile data from LinkedIn, but more online accounts are being targeted. The technology for syntactic matching will be further improved through linguistic analysis. Our most challenging future enhancement is the envisaged semantic extension to the current syntactic-based profile attribute matching. Research contributions will on the other hand focus on defining an appropriate semantic-based attribute weighting for each matched attribute, together with the definition of a metric which takes into account all the resulting weighted matches and

the identification of a threshold that determines whether two or more online profile refer to the same person. Online posts are also taken into consideration [24]. An analysis of posts from multiple accounts can help us discover whether two or more online profiles are semantically equivalent.

Finally, a comprehensive evaluation of our system would be performed on three levels — i) syntactic matching, ii) semantic matching, and iii) a combination of. This would help determine whether our part-syntactic, part-semantic approach actually yields better results.

Acknowledgements. 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*).

References

1. W. Akhtar, J. Kopecky, T. Krennwallner, and A. Polleres. Xsparql: Traveling between the xml and rdf worlds and avoiding the xslt pilgrimage. In *Proc. 5th European Semantic Web Conference (ESWC2008)*, pages 432–447, Berlin, Heidelberg, 2008.
2. D. Appelquist, D. Brickley, M. Carvahlo, R. Iannella, A. Passant, C. Perey, and H. Story. A standards-based, open and privacy-aware social web. W3c incubator group report, W3C, december 2010.
3. D. Aumueller, H. Do, S. Massmann, and E. Rahm. Schema and ontology matching with coma++. In *Proc. ACM SIGMOD international conference on Management of data*, pages 906–908, New York, NY, USA, 2005.
4. D. Berrueta, D. Brickley, S. Decker, S. Fernndez, C. Grn, A. Harth, T. Heath, K. Idehen, K. Kjernsmo, A. Miles, A. Passant, A. Pollares, and L. Polo. Sioc core ontology specification. Technical report, 2010.
5. S. Bortoli, H. Stoermer, P. Bouquet, and H. Wache. Foaf-o-matic - solving the identity problem in the foaf network. In *Proc. Fourth Italian Semantic Web Workshop (SWAP2007)*, 2007.
6. D. Brickley and L. Miller. Foaf vocabulary specification 0.98. Technical report, 2010.
7. A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Proc. Workshop on Wordnet and other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001.
8. W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proc. IJCAI-03 Workshop on Information Integration (IIWeb-03)*, pages 73–78, Acapulco, Mexico, Aug.9-10 2003.
9. V. Cross. Fuzzy semantic distance measures between ontological concepts. In *Proc. Fuzzy Information, 2004. Processing NAFIPS '04. IEEE Annual Meeting of the*, volume 2, pages 635–640, june 2004.
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. E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. Twentieth International Joint Conference for Artificial Intelligence (IJCAI-07)*, pages 1606–1611, Hyderabad, India, Jan.6-12 2007.

12. J. Golbeck and M. Rothstein. Linking social networks on the web with foaf: A semantic web case study. In *Proc. Twenty-Third Conference on Artificial Intelligence (AAAI'08)*, pages 1138–1143, Chicago, Illinois, USA, 13-17 2008.
13. M. Ion, L. Telesca, F. Botto, and H. Koshutanski. An open distributed identity and trust management approach for digital community ecosystems. In *Proc. International Workshop on ICT for Business Clusters in Emerging Markets, June 2007. Michigan State University*, 2007.
14. N. Jian, W. Hu, G. Cheng, and Y. Qu. Falcon-ao: Aligning ontologies with falcon. In *Proc. K-Cap 2005 Workshop on Integrating Ontologies. (2005)*, pages 87–93, 2005.
15. S. Labitzke, I. Taranu, and H. Hartenstein. What your friends tell others about you: Low cost linkability of social network profiles. In *Proc. 5th International ACM Workshop on Social Network Mining and Analysis*, San Diego, CA, USA, Aug.20 2011.
16. Y. Li, Z. A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15:871–882, 2003.
17. P. Mika. Flink: Semantic web technology for the extraction and analysis of social networks. *Journal of Web Semantics*, 3(2-3):211–223, 2005.
18. A. Monge and C. Elkan. The field matching problem: Algorithms and applications. In *Proc. Second International Conference on Knowledge Discovery and Data Mining*, pages 267–270, 1996.
19. A. Mylka, L. Sauermann, M. Sintek, and L. van Elst. Nepomuk contact ontology. Technical report, 2007.
20. E. Raad, R. Chbeir, and A. Dipanda. User profile matching in social networks. In *Proc. 13th International Conference on Network-Based Information Systems, 2010*, pages 297–304, Takayama, Gifu Japan, 2010.
21. S. R. Ray. Interoperability standards in the semantic web. *Journal of Computing and Information Science in Engineering, ASME*, 2:65–69, 2002.
22. M. Rowe and F. Ciravegna. Getting to me: Exporting semantic social network from facebook. In *Proc. Social Data on the Web Workshop, International Semantic Web Conference*, 2008.
23. L. Sauermann, L. van Elst, and K. Mller. Personal information model (pimo). Oscaf recommendation, OSCAF, february 2009.
24. S. Scerri, K. Cortis, I. Rivera, and S. Handschuh. Knowledge discovery in distributed social web sharing activities. In *Making Sense of Microposts (#MSM2012)*, pages 26–33, 2012.
25. S. Scerri, R. Gimenez, F. Herman, M. Bourimi, and S. Thiel. digital.me - towards an integrated personal information sphere. In *Proc. Federated Social Web Europe Conference (FSW 2011)*, Berlin, Germany, 2011.
26. P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. *Journal on Data Semantics IV*, 3730:146–171, 2005.
27. S. A. Takale and S. S. Nandgaonkar. Measuring semantic similarity between words using web documents. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 1(4), 2010.
28. H. Yang and J. Callan. Learning the distance metric in a personal ontology. In *Proc. 2nd international workshop on Ontologies and information systems for the semantic web*, pages 17–24, New York, NY, USA, 2008.