Looking for Experts? What can Linked Data do for You?

Milan Stankovic
Hypios & STIH, Université Paris-Sorbonne
187 rue du Temple, 75003 Paris, France
+33 6 74 22 57 12
milan.stankovic@hypios.com

Claudia Wagner
JOANNEUM RESEARCH
Steyrergasse 14, 8010 Graz
Austria
+43 316 876 2617
claudia.wagner@joanneum.at

Jelena Jovanovic
FON, University of Belgrade
Jove Ilica 154, 11000 Belgrade, Serbia
+381 11 3950 853
jeljov@fon.rs

Philippe Laublet
STIH, Université Paris-Sorbonne
28 rue Serpente, 75006 Paris, France
+33 1 53 10 58 25
philippe.laublet@paris-sorbonne.fr

ABSTRACT

Expert search and profiling systems aim to identify candidate experts and rank them with respect to their estimated expertise on a given topic, using available evidence. Traditional expert search and profiling systems exploit structured data from closed systems (e.g. email program) or unstructured data from open systems (e.g. the Web). However, on today's Web, there is a growing number of data sets published according to the Linked Data principals, the majority of them being part of the Linked Open Data (LOD) cloud. As LOD connects data and people across different platforms in a meaningful way, one can assume that expert search and profiling systems would benefit from harnessing LOD. The work presented in this paper sets out to prove this assumption and to explore potential benefits and drawbacks of using the LOD cloud as expertise evidence source. We conducted several experiments to evaluate the feasibility of existing expert search and profiling approaches on a recent snapshot of the LOD cloud. Our findings indicate that LOD cloud is already a useful source for some kinds of expert search approaches (e.g., those based on publications and professional events) but still has to meet certain requirements in order to reach its full potential.

1. INTRODUCTION

Expert Finder systems are Information Retrieval (IR) systems which identify candidate experts and rank them with respect to their estimated expertise on a given topic, using available evidence (e.g. documents about/of candidates, social networks of candidates, activities of candidates in real world and online). In literature, expertise is often defined as 'high, outstanding, and exceptional performance which is domain-specific, stable over time, and related to experience and practice' [1]. The nature of expertise itself as well as the fact that people grow and change over time, make solving expert finding and profiling difficult [2]. Accordingly, expert profiling and search have been quite extensively covered research topics, with lots of research efforts directed towards identifying experts, especially within the organizational context.

Traditional expert search and profiling systems exploit either structured data from closed systems (e.g. email program) or unstructured data from open systems (e.g. the Web). The former approaches tend to follow a 'closed-world' view and make inferences about people's expertise based on evidences collected from a closed system (e.g., a repository of scientific publications, messages exchanged in Q&A forums, etc.). What these approaches lack is a comprehensive, 'open-world' view of a person's expertise based on evidences that originate from diverse, and often distributed sources on the Web (e.g., person's CV, professional online and offline activities, his/her social network, etc). The later approaches follow an 'open-world' view, but suffer from limited inference mechanism due to the use of unstructured data. What these approaches lack is a comprehensive understanding of the meaning of the data and the relations amongst them.

However, on today's Web, there is a growing number of data published according to the Linked Data principals[1], the majority of them as a part of the Linked Open Data (LOD)[2] cloud. Thanks to the property of being interlinked, this emerging mass of data might, be a promising source for expert search. In this paper we take the challenge of analyzing the potentials and drawbacks of the currently available datasets in the LOD cloud for the expert retrieval and profiling task. We explore if the assumptions about what makes an expert (so-called expertise hypotheses) taken by traditional approaches can be used for expert finding on the current LOD cloud. Furthermore, we investigate if LOD (and Linked Data in general) can open possibilities for novel expertise hypotheses and try to unveil the advantages of LOD over traditional expert search approaches. Finally, we give recommendations for what needs to be done to make LOD and Linked Data in general, an even better source for expert search.

The reminder of this paper is organized as follows: in Section 2 we review expert profiling and search approaches from literature and distill their core assumptions, so-called expertise hypotheses.

[1] http://www.w3.org/DesignIssues/LinkedData.html

[2] In this paper we use the term "Linked Data" to refer to the publishing principals, and "LOD cloud" to refer to the interlinked, publicly accessible datasets published using those principals and available at: http://richard.cyganiak.de/2007/10/lod/.

In Section 3 we describe how we investigated the feasibility of different expertise hypotheses on the LOD sources and share the results of our empirical analysis. Subsequently, in Sections 4 and 5 we report both potentials and pitfalls, which we noticed during our study of using the LOD cloud as expertise evidence source. We conclude this work by suggesting the directions for future work that would make LOD and Linked Data in general, even more useful for determining who knows what on the Web.

2. EXPERTISE HYPOTHESES

In the existing literature on expert finding, different authors make different assumptions on what makes an expert and how expertise can be assessed. We call these assumptions *expertise hypotheses*. In general, an expertise hypothesis can be interpreted as a rule containing a condition and a conclusion that a particular user is an expert in specific domain of competence:

If (condition) then user A might be an expert in the domain X.

The *condition* involves a mention of *user A*, the *domain of expertise X*, and a binding element that allows for connecting the two. This binding element is what we call the *evidence of competence*. For example, a book that a user wrote about Quantum Physics might be an evidence of his/her expertise in that domain. Expertise hypotheses are thus the key assumptions of every expert search approach. Accordingly, we use these hypotheses as abstractions of expert search approaches in our effort to evaluate their feasibility with datasets of the LOD cloud.

Faced with different data, different context of expert search and different goals of their projects, different authors have adopted different expertise hypotheses. In this chapter we investigate the nature of those hypotheses and offer their classification. The classification of expertise hypotheses should bring different facets of expertise to light and give ground for understanding how one can go from raw data to expertise assessment. This understanding will be essential for the evaluation of LOD potentials and for deriving the requirements for LOD-based expert finding approaches.

2.1 Classification and Review of selected Expertise Hypotheses

Nowadays, in times of the Social Web, users tend to leave their traces on the Web. These traces can serve as evidences of users' expertise. Different expert finding approaches use different types of expertise hypothesis which rely on different types of evidence data. Some rely on the content a user created/collected/shared; others on reliable sources of information about a user (e.g., Wikipedia) and so on. In general, we can distinguish among 3 major kinds of hypotheses based on the type of evidence they rely upon:

- hypotheses that rely on content that is related to an expert candidate;
- hypotheses that rely on activities of an expert candidate;
- hypotheses that rely on the reputation and authority of an expert candidate.

Content-based hypotheses take into account the content that a user has created and/or the content a user owns. Hypotheses related to activities take into account either a user's online activities, or the activities that a user performs in the offline world. The third type of hypotheses takes into account the opinions of other users about a given user and a user's social network.

In the following section we present a selection of expertise hypotheses that we found in literature as well as those that we think might be of interest for the future expert finding approaches that make use of the LOD cloud. Among many hypotheses that we have found, we have chosen 18 that we find most compelling for expert search on today's Web. We favored the diversity of expertise evidences in the selection of hypotheses.

2.1.1 Hypotheses related to user's online content

In this section we present different hypotheses that are related to a user's online content. In literature we found a number of hypotheses that deal with content created by users and content owned by users (where the former is more often used and considered more important). Since the information about content owned by users (e-mails [3], documents, scientific articles, etc.) is mostly not available on the Web, we focus here only on hypotheses that deal with content created by users.

H1: *If a user wrote a scientific publication on topic X than he might be an expert on topic X*

In many approaches, scientific publications are used to identify experts in a certain field. This hypothesis is quite convenient because peer-review of scientific publications guaranties the relevance and quality of authors' writings. However, the expertise level of a user may also depend on the impact of the journal or conference where the paper was published and the number of papers a user published. In addition, it is not always easy to relate the authors of a paper with the domains of expertise that the paper identifies. In [4] a simple lexical pattern-matching approach is used to identify topics of a paper and then assume the expertise of paper authors for those topics. Demartini & Niederée [5] use Semantic Desktop to identify experts. They suppose a scenario when a desktop user needs to ask a domain-related question, and the system then searches for experts in the given domain by leveraging the content stored on the user's computer. Their approach takes scientific papers available on the user's computer and ranks all the authors it can find. Although this approach uses other data as well (e-mails, PDF and DOC files, etc.), it takes a rather closed-world view, as it cuts the user's computer of the outside world. The resulting expert ranking is highly sensible to the data that the user possess and would benefit from the possibility to include external data into calculation.

H2: *If a user wrote a Wikipedia page on topic X than he might be an expert on topic X.*

Wikipedia has grown a dedicated community of moderators who make sure all the content is backed up with references, and that reliable content is not replaced by manipulative users. Having contributed a reliable content to a Wikipedia page indicates that the contributor is knowledgeable on the topic of the page. One of the approaches that take advantage of Wikipedia to find experts is presented in [6]. Once the experts are identified, various techniques are used to rank them, including a PageRank-like and HITS-like algorithms which is applied on the link structure of Wikipedia articles in order to identify the most influential pages (and their authors).

H3: *If a user blogs a lot about topic X, then he might be an expert for topic X*

Several approaches exist which exploit the blogosphere as expertise evidence source. For example, Kolari et al. [7] rely on internal corporate blogs to find experts inside a particular company, IBM. However, their approach can easily be

generalized to the blogs on the Web. A similar approach is taken by [8] and [9].

2.1.2 Hypotheses related to user's activities

In this section we present hypotheses related to users' activities. We distinguish between online and offline activities.

2.1.2.1 Hypotheses related to user's online activities

This section presents hypotheses which assume that a user's online activities related to a certain topic imply his/her expertise in that topic.

H4: *If a user answers questions (on topic X) from experts on topic X then he might himself be an expert on topic X.*

This hypothesis is mostly used in approaches that rely on Questions & Answers (Q&A) communities. For instance, the work presented in [10] uses Yahoo! Answers[3] community to identify experts.

This hypothesis can also be useful in an alternated form that would take into account the level of expertise in order to rank the expert candidates. In that sense, the level of expertise of a user who answers a question might be evaluated as a function of the level of the user that posed it. Jurczyk and Agichtein [11] use a sophisticated approach based on link analysis to identify experts in Q&A communities. They construct a graph out of users' interactions in the social network: when a user A answers a question of a user B, a connection from B to A is created. The resulting graph can then be exploited by PageRank-like and HITS-like algorithms in order to propagate the expertise through the graph and select the best experts. The rank of the user who posted a question is influencing the gain in rank of users who post answers. A similar approach is taken in [12], where Java support forum is used as a source of questions and answers.

H5: *If a user is among the first to discover (and share) important resources (i.e. resources which become later popular) on topic X, then he might be an expert on topic X.*

Noll et al. [13] use bookmarks that users save online, as identifiers of expertise. They consider a user's ability to find good Web resources on a particular topic (and save them as bookmarks) to be a proof of a user's expertise. The fact that a Web resource is later endorsed by many users makes it possible to conclude that it is a high-quality Web resource. The authors especially focus on the time of bookmarking and consider those users that are the first who find and share a good resources as experts.

H6: *If a user participates in collaborative software development project then he might be an expert in the programming language that is used in the project.*

Although we haven't found any literature that would describe such an approach, we believe that with the growing number of online communities for code sharing and collaborative coding, software development projects[4] might be a good evidence of programming expertise.

2.1.2.2 Hypotheses related to a user's real life activities and achievements.

In this section we present hypotheses related to activities that a user performs in real life (but that may as well be traced online).

H7 *If a user claims in his resume/CV that he is skilled in a topic X than he might be expert in topic X.*

On their homepages, online CVs, as well as user profiles in online communities, people tend to claim that they have particular skills. Although we have not found an expert search approach that is based purely on these data, we found it a useful source for expert mining.

H8: *If a user has obtained funded research grants in a certain (domain) field, then he might be an expert in that field.*

SAGE (Searchable Answer Generated Environment) Expert Finder, which serves as a searchable repository of experts in Florida universities, was developed on the premise that researchers who successfully obtain funded research grants are experts in their fields [14]. Even though widely recognized, this is not a perfect indicator of expertise, because the available data (on funded projects) do not provide the granularity that would be required to identify the level of expertise. In addition, SAGE Expert Finder acts in a closed environment, as it has access only to the data about the funded projects of Florida universities.

For the hypotheses H9 to H16 we haven't found previous research that made use of these hypotheses, but we found them relevant and applicable using the (semi-structured) data of professional social networks (e.g., LinkedIn[5]), personal homepages, and homepages of professional events.

H9: *If a user has a certain position in company then he might be an expert on the topic related to his position.*

H10: *If a user supervises/teaches someone then he might be an expert on the topic he/she teaches.*

H11: *If a user has several years of experience with working on something related to topic X then he might be an expert in topic X.*

H12: *If a user is a member of the organization committee of a professional event, then he might be expert on the topic of the event.*

H13: *If a user is giving a keynote or invited talk at a professional event, then he can be considered an expert in the domain topic of the event.*

H14: *If a user is a chair of a session within a professional event, then he can be considered an expert in the topic of the session (and by generalization, also an expert in the domain topic of the event).*

H15: *If a user is presenting within a session of a professional event, then he can be considered an expert in the topic his presentation is about. By generalizing, he can be considered an expert in the topic of the session/event his presentation is part of.*

H16: *If a user was an invited guest on a show (published on the Web as a podcast and/or video streaming) on the topic X, then he might be an expert in the topic X.*

2.1.3 Hypotheses related to a user's reputation and authority

The hypotheses presented in this section do not take into account information produced by expert candidates, but information about them, i.e. their reputation or perceived authority.

[3] http://answers.yahoo.com/

[4] For instance http://sourceforge.net and http://code.google.com

[5] http://linkedIn.com

H17: *If a user's blog about a topic X gets lost of comments, then he might be an expert for topic X.*

The approach by Kolari et al. [7] that we have already discussed also uses this hypothesis in addition to H3.

H18: *If a user has high social connectedness with an expert in topic X, then he is considered to be an expert in topic X.*

This hypothesis is used to propagate expertise within a network of users. It is especially useful when an initial (seed) set of experts in the community is already known. This is the case in [15], where social connectedness is calculated based on e-mails and documents that relate two users.

3. Experiments

In this section we present the experiments that we have conducted to evaluate selected expertise hypotheses on the LOD cloud. The aim of our experiments was (1) to find out if and how certain expertise hypotheses can be evaluated based on the LOD cloud as source for expertise evidence and (2) to explore whether LOD (and Linked Data in general) has advantages over traditional approaches. For practical reasons we relied on Richard Cyganiak's version of LOD cloud made on 14.07.2009[6] (the latest one at the time of our evaluation). We thus apologize to the maintainers of all the datasets that appeared in the meantime, whose efforts we could not take into account. We also rely on our map of Linked Data Related to Competence[7]. This map helps to identify the data sources in the LOD Cloud that contain data about different kinds of evidence of expertise.

3.1 Experimental Setup

In order to be a useful evidence source for expert search, the LOD cloud needs to satisfy certain conditions. We have designed the following tests to verify if those conditions are met by the current LOD cloud and conducted these test for each particular expertise hypothesis.

Test 1: *Does the LOD cloud contain datasets with the type of data that is needed for expert search using a particular expertise hypothesis?*

The first test verifies if the LOD cloud contains a dataset that claims to provide the kind of data required for a particular expertise hypothesis. For example, for a hypothesis which uses scientific articles to evaluate expertise, LOD passes this test if it contains a dataset providing data about scientific papers.

Test 2: *Do relevant data sources of the LOD cloud contain all data which are necessary to evaluate a particular expertise hypothesis?*

The second test shows if the LOD sources that are relevant for a particular hypothesis, expose their data with the necessary level of detail. For example, an expertise hypothesis might take into account the time of saving a bookmark. If a respective data source about bookmarks would not contain bookmarking date-time data, it would be useless for expert finding approaches using this hypothesis.

Test 3: *Does the LOD cloud contain links between data sources that are necessary to identify domains of expertise?*

The third test verifies whether for a specific hypothesis, relevant LOD sources contain links that allow establishing a connection between a user and his domain of expertise. This connection is usually established through the evidence of expertise that needs to point to a certain topic of expertise.

This test shows how easily results of LOD-based expert search can be combined with Web systems that use semantic annotations (e.g. semantic tagging systems, semantic microblogging, etc.). For instance, recommender systems, content personalization etc. might be possible ways to mash up expert finding and other Web systems.

Test 4: *Does LOD cloud contain links between user data belonging to the same real world person?*

Apart from being able to connect evidence data with domains of expertise, for some advanced scenarios it is necessary to integrate data about a given user from different data sources. The fourth test proves whether for a specific hypothesis relevant LOD sources contain links which connect distributed user identities.

This test shows if an approach based on a particular hypothesis can easily combine data about a user from several sources. This would allow systems to infer the expertise of a user based on a more comprehensive set of data about a user.

These tests are performed using several techniques of examining the LOD cloud. First, in Test 1, we have used the existing information about the datasets in order to find relevant LOD datasets for particular types of hypotheses. For most datasets there is a description of its content on the dataset's homepage, as well as an example URI that helped us to get a general insight into the dataset's content. If we could find a dataset claiming to contain the required type of data, we noted a positive mark (plus sign in Table 1).

In order to verify if datasets contain the relevant data for evaluating a given expertise hypothesis (e.g., in case of H14 we searched for data about participants' roles in the SW Conference data set), we used Sindice[8] to search for the use of relevant classes and properties in the cloud and thus see what data is present. In addition to Sindice, we also used available SPARQL[9] endpoints providing access to LOD datasets and ran simple DESCRIBE queries in order to get full descriptions of relevant resources. As the final step we ran SPARQL queries on endpoints to check the existence of relevant properties and their values in the dataset. Only if we obtained no results in any of the three steps, we noted a negative mark (minus sign in Table 1).

We conducted Test 3 and 4 in a similar way (using Sindice, and then querying the SPARQL endpoints) as Test 2. In some cases, where several data sets were relevant for a hypothesis, we had to note a neutral mark ('+-'). The neutral mark indicates that the usage of the hypothesis would be possible on the current LOD cloud, but not all the relevant sources would reply properly. This situation also occurs when a dataset fulfills the test partially (by providing data/links only for a portion of its data), or in cases where we have several data sets that offer different data richness.

[6] http://richard.cyganiak.de/2007/10/lod/

[7] http://research.hypios.com/mstankovic/lod-competence/

[8] http://sindice.com is a Semantic Web Index and Search Engine

[9] http://www.w3.org/TR/rdf-sparql-query/

Hypothesis type	Hypothesis	Are there relevant datasets?	Are there useful facts in the datasets?	Are there necessary links to categories?	Are there necessary links to connect user data?
Authorship of High-quality Content	H1: If a user wrote a scientific publication on topic X than he might be an expert on topic X	+	+	+-	+
	H2: If a user wrote a Wikipedia page on topic X than he might be an expert on topic X.	+	+	+	-
	H3: If a user blogs a lot about topic X, then he might be an expert for topic X	+	+	+-	+-
Online Activities	H4: If a user answers questions (on topic X) from experts on topic X then he might himself be an expert on topic X	+	-	-	-
	H5: If a user is among the first to discover (and share) "important/good" resources (i.e. resources which become later popular) on topic X, then he might be an expert on topic X.	+	-	+	-
	H6: If a user participates in collaborative software development project then he might be an expert in the programming language that is used in the project.	+	+	+-	+-
Real Life Activities and Achievements	H7 If a user claims in his resume/CV that he is skilled in a topic X than he might be expert in topic X.	-	-	-	-
	H8: If a user has obtained funded research grants in a certain (domain) field, then he might be an expert in that field.	+	+	-	+
	H9: If a user has a certain position in company then he might be an expert on the topic related to his position.	+	-	-	+-
	H10: If a user supervises/teaches someone then he might be an expert on the topic he/she teaches.	-	-	-	-
	H11: If a user has several years of experience with working on something related to topic X then he might be an expert in topic X.	-	-	-	-
	H12: If a user is a member of the organization committee of a professional event, then he might be expert on the topic of the event.	+	+	-	+
	H13: If a user is giving a keynote or invited talk at a professional event, then he can be considered an expert in the domain topic of the event.	+	+	-	+
	H14: If a user is a chair of a session within a professional event, then he can be considered an expert in the topic of the session (and by generalization, also an expert in the domain topic of the event).	+	+	-	+
	H15: If a user is presenting within a session of a professional event, then he can be considered an expert in the topic his presentation is about. By generalizing, he can be considered an expert in the topic of the session/event his presentation is part of.	+	+	-	+
	H16: If a user was an invited guest on a show (published on the Web as a podcast and/or video streaming) on the topic X, then he might be an expert in the topic X.	-	-	-	-
Reputation and Authority	H17: If a user's blog about a topic X gets lost of comments, then he might be an expert for topic X.	+	+	+-	+-
	H18: If a user has higher social connectedness with an expert in topic X, then he is considered to be a better expert in topic X	+	+	+-	+-

Table 1 The evaluation results by hypotheses

For some data sets (e.g. SIOC sites) there was no unique SPARQL endpoint, so we relied on Sindice to examine the existing data, and we also tried to find relevant data exporters[10] (and sites that use them[11]) to verify which data they provide. With regard to the data we found, we gave positive, negative or neutral marks.

3.2 Results

In this section we present the results of our evaluation of the feasibility of the expert search on LOD cloud. Table 1 gives a summary of the results. For each hypothesis we applied all four tests on the LOD datasets.

For the first hypothesis - H1 we found many LOD data sources that contain useful data about scientific publications (Table 2). Those datasets are interlinked among each other and allow for easy data merging and finding the publications of one author across various datasets. All the data that an expert search approach based on H1 might need is present. However, links to semantic descriptions of keywords and categories of scientific

[10] e.g., http://sioc-project.org/exporters

[11] Wherever the list of sites that use an exporter was available.

papers are often missing; apart from the DBLP Berlin dataset that has links to DBPedia[12]. Another (positive) exception is the SW Conference Corpus dataset (storing data about publications on Semantic Web conferences) that is also rich in topics from DBPedia, as well as in inverse functional properties. Thus it can already be used for H1-based approaches. For instance, the query presented in Figure 1, run on this dataset gives meaningful results (the known Semantic Web researchers – authors of Semantic Web papers).

The hypothesis H2 is relatively well-covered thanks to the SIOC MediaWiki Exporter[13] that exports the data about the authors (contributors) of Wikipedia articles. The articles themselves represent an identification of expertise domains, but the nature of user data in Wikipedia does not make it easy for SIOC MediaWiki Exporter to expose the unique identifiers for the content authors and interlink data about them from elsewhere.

SIOC sites dataset is a valuable source of blog-related data needed for H3. The key to usefulness of SIOC data for expert search is the availability of topic information. Using Sindice, one can find many SIOC sites that provide such information using the sioc:topic property. But there are also many SIOC sites that do not provide data about their topics by interlinking them with semantic concepts denoting the meaning of topics. Recent Approaches for Semantic Tagging (like CommonTag[14]) could help bridge this gap by augmenting blog posts with crowd-sourced categories that make a reference to DBPedia concepts.

```
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX akt: <http://www.aktors.org/ontology/portal#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX swrc: <http://swrc.ontoware.org/ontology#>
PREFIX swc: <http://data.semanticweb.org/ns/swc#>

SELECT DISTINCT $person
WHERE {
   {$person a foaf:Person} UNION {$person a akt:Person}.

    { $paper swrc:author $person} UNION { $paper dc:creator
$person} UNION { $paper foaf:maker $person} UNION  { $paper
akt:has-author $person}.

  { $paper swc:hasTopic dbpedia:Semantic_Web} UNION { $paper
sioc:topic dbpedia:Semantic_Web} UNION { $paper
dcterms:subject dbpedia:Semantic_Web}

}
```

Figure 1 SPARQL query for finding experts using H1

For approaches based on H6, a useful data source is RDFOhloh – the export of data related to software development projects that take place at Ohloh[15]. This source provides both inverse functional properties for the members of the projects, and links to DBPedia concepts identifying the programming languages that are used. It is thus perfectly suited for finding experts on specific programming languages.

H12 – H15 are related to professional events. At present, SW Conference Corpus dataset provides this kind of data for Semantic Web-related professional events. We hope that events from other domains will be represented in a similar way in a near future. Data about topics of events are mostly missing. However, a workaround is possible, since the papers presented on events are usually annotated with topics, which may help infer the topic of the event in general. Via the assumption that the topics of a professional event is a union of all the topics associated with papers presented on the event, we can get the list of people that had a certain role on a Semantic Web-related event. The query shown on the Figure 2 gives the top Semantic Web researchers who were chairs of events related to this domain, thus proving that H14 is fully feasible on the present LOD cloud.

This dataset is also rich in inverse functional properties that allow identifying a user in other data sets and merging the user data across datasets.

```
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX akt: <http://www.aktors.org/ontology/portal#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX swrc: <http://swrc.ontoware.org/ontology#>
PREFIX swc: <http://data.semanticweb.org/ns/swc#>

SELECT DISTINCT $person

WHERE {
   {$person a foaf:Person} UNION {$person a akt:Person}.

    $person swc:holdsRole $role.
    $role swc:isRoleAt $event.
    $role a swc:Chair.
    $event swc:hasRelatedDocument $proceedings.
    $paper swc:isPartOf $proceedings.

{ $paper swc:hasTopic dbpedia:Semantic_Web} UNION
{ $paper sioc:topic dbpedia:Semantic_Web} UNION
{ $paper dcterms:subject dbpedia:Semantic_Web}

}
```

Figure 2 SPARQL query for finding experts using H14

H18-based approaches can already significantly benefit from the LOD cloud, as social connectedness can be evaluated through FOAF files that disclose connections between users, but also trough SIOC sites that contain traces of users interactions that can serve to measure the extent of connectedness.

Table 2 gives a summary of all the cases where the current LOD cloud contains the data relevant to a certain hypothesis. Some of those positive cases were detailed in this section. Section 4 gives an overview of advantages that the use of LOD has for Expert Search.

Although the current LOD is already a useful source for expert search, it still has to advance to allow for deducing expertise based on further hypotheses. More data sets are needed to make H4, H7, H10, H11, and H16 feasible. H5 and H9 would benefit from more (detailed) data in the existing datasets; and H8 and H17 would benefit from new links between data sets. Section 5 considers those pitfalls of the current LOD cloud in more details.

[12] http://dbpedia.org

[13] http://ws.sioc-project.org/mediawiki/

[14] http://www.commontag.org/

[15] http://www.ohloh.net/

Hypothesis	LOD Data Sets that Contain the Evidences of Competence[16]
H1	SemanticWeb.org; SW-Conference Corpus; ECS Sauthampton; LAAS-CNRS; CiteSeer; IBM; Pisa; IEEE; ACM; RKB ECS Southampton; eprints; IRIT Toulouse; Newcastle; RAE 2001; Budapest BME; DBLP RKB Explorer; DBLP Hannover; DBLP Berlin
H2	SIOC sites (SIOC wiki)
H3	SIOC sites
H5	Faviki, Virtuoso (via Sponger)
H6	DOAP Store, RDFOhloh
H8	Cordis, National Science Foundation
H9	ChrunchBase
H12-H15	SW Cofnerence
H17	SIOC sites
H18	FOAF profiles, SIOC sites

Table 2 Relevant data sets for some hypotheses

4. POTENTIALS

Using Linked Open Data for expert search has various advantages over the traditional approaches that use unstructured data. In this section we discuss some of those advantages, we observed during our analysis presented in Section 3.

4.1 Decoupling Data from Hypotheses

Most of the standard, non-Linked Data based, approaches for expert finding deal with a corpus of non structured data. They process the data using some kind of Information Extraction technique and try to extract valuable traces for expert identification.

However, the particular way in which those standard approaches extract structured data from their heterogeneous data corpuses is often inspired by the expertise hypothesis in use. Once the corpus is treated, the extracted data are stored in a data structure that fits a particular hypothesis. As an example, an expert search approach might search for experts among authors of academic journals. Thus, it would extract the triples <expert, journal paper, topic> from the journal corpus. These data would then be useless for a different approach that considers early adopters of a topic as more valuable experts, because it needs the data about the time of publications.

On the other hand, in the Linked Data based approach the expertise hypothesis and the data structure are decoupled. The data in the LOD cloud are not supposed to be tailored for any specific expertise hypothesis/approach. Instead it is provided in a form that supports multiple purposes. The expert search approaches built on top of LOD cloud thus provide a higher degree of flexibility and adaptability. The same Web data can be useful for finding experts in many domains, and in many different ways.

4.2 Unlimited, Cross-Platform Evidence

Traditional expert search systems usually exploit only a limited set of platforms as source for expertise evidence data, because they need to 'understand' the data schema of different data sets and need to know how to combine them in order to apply expertise hypothesis on them. Linked Data based expertise systems have the power to overcome this limitation by exploiting the whole Linked Data sphere to search for expertise evidence. That means, general expertise hypothesis (such as, *a user is an expert if he publishes high quality content about a topic*) can be applied to various data sets stemming from various platforms. Widely-used vocabularies for describing datasets and data itself create a common data schema layer and allow expert search systems to access an open and distributed set of data sources. Links between different datasets identify relations between data items. For example, equivalence relations allow for identifying equivalent items in different data sets (e.g. user accounts belonging to the same real life person or product descriptions about the same real world thing).

Expert search systems can obviously benefit from harnessing distributed, interlinked data, because they obtain a completer picture of an expert candidate, his activities, content and social network. Furthermore, by harnessing distributed comments/opinions/ratings about the content produced by an expert candidate, Linked Data based expert search systems can use a greater variety of opinions to estimate the quality of an expert candidate's content and his authority and reputation. Finally, besides wanting to know whether a person who can answer their queries or meets their criteria exists, seekers of experts also want to know how extensive the expert's knowledge or experience is, whether there are other persons who could serve the same purpose, how he/she compares with others in the field, how the person can be accessed (contacted), etc. [16]. So, besides expert profiling, there are additional requirements that have to be addressed for a fully fledged expert finder system. By leveraging the 'linking' aspect of Linked Data and the ability to navigate through and integrate disparate datasets, one would be better able to address all these questions and requirements than it would have been without the linking effect.

5. PITFALLS

Despite the benefits and potentials Linked Data has for expert search, expert search system developer and researcher must also consider existing pitfalls when using the currently available LOD cloud as expertise evidence source. In this section we present the problems that result from our analysis (based on the test cases presented in Section 3).

5.1 Usage Restricted Data

In some cases the expert search relies on data that is inherently private in nature and cannot be used by everyone (e.g. e-mails and similar personal content, as well as the majority of content in corporate intranets). Such data are usually not linked with the rest of the Web's data. Thus the approaches based on e-mails, private documents, attention records, intranet documents, etc. do not work with the current LOD cloud. However, the present state of things is not a fault of Linked Data itself, but rather of the lack of the implemented authorization mechanisms, that might work with Linked Data. The fact that e-mails are private does not mean that they cannot be made available as Linked Data (possibly

[16] The names of data sets correspond to the names used on the LOD cloud diagram. We refer the readers interested in homepages of those data sources to the clickable version of the LOD diagram available at: http://richard.cyganiak.de/2007/10/lod/

interlinked with FOAF user profiles and DBPedia concepts) to those who should have the access to them.

Existing security mechanisms, such as OAuth[17] and FOAF+SSL [17], allow protecting private data and metadata even if they are published as Linked Data. However, no significant amount of private and/or usage-restricted Linked Data has been published yet.

There is also a lack of motivation for publishing private (individual and corporate) data as Linked Data. Being aware of the general lack of understanding the benefits offered by publishing data as Linked Data, the Linked Data community has recently started exploring business models for publishing and consuming Linked Data [18], [19]. We hope that these efforts will result in better understanding of the difference between Linked Data and Linked Open Data (presently often mistakenly considered the same [20]), as well as in gradual, but steady increase in (personal and organizational) private data exposed as Linked Data.

5.2 Lack of Data

In some cases the current LOD cloud is not a good source for expert finding because it simply does not contain the kind of data needed for a certain hypothesis. During our evaluation we have identified the kinds of data that would be a useful source for expertise evidence, but are missing in the current LOD cloud. Examples of data that the LOD cloud might benefit from are presented in the reminder of the section.

Q&A sites are a useful source of data about expertise, and despite the possibility to represent them using the SIOC ontology, we have not found any such website that provides SIOC-based data export. H4 is thus not applicable on the current LOD cloud.

Data about careers of people is just another example of data that is lacking. There are no good reasons why data about university diplomas and jobs would not be in LOD or otherwise linked with LOD. In fact having it would make it easy to verify the claims of professional achievements. The trend of making data public is obvious (e.g., USA government initiative[18], UK government initiative[19]). Therefore, we expect that university and corporate structure data become a part of the LOD cloud. Approaches based on H7, H10 and H11 would benefit from these data.

Professional podcasts with guest experts[20], video lectures[21], as well as online slide presentations[22] would have been a valuable data source for expert profiling if the data about the hosted resources and their authors were available in RDF (especially for H16).

Public mailing lists are a valuable source of expertise-related data. However we have not found many mailing lists, which expose their data in RDF. The project SWAML[23] provides an SIOC-based exporter for mailing lists that can be used for exporting the public data from these lists.

Data about professional events is for now only present for Semantic Web-related events in the SW Conference dataset, but many other professional events from other domains stay unrepresented in LOD cloud.

User activities, like attending the professional events, giving presentations, etc. are also lacking in the current LOD cloud. Although those data become more and more public though the emergence of Twitter[24], and the more liberal privacy settings on Facebook[25], they are not presented in structured form and are consequently not part of the LOD cloud.

We hope that our identification of useful data sources and the ontologies that might be used for data publishing might inspire some future work on making that data available as Linked Data. The ExpertFinder initiative[26] also provides some vocabularies for exposing expertise–related data as well as the incentives for publishing it.

5.3 Lack of Details

In the case of some hypotheses, the necessary kinds of data exist, but metadata descriptions are not fine-grained enough and details needed by the expertise hypothesis are missing.

Faviki[27] is a good example of this issue as well. It provides useful data about tagging with links to DBPedia, but the data about the time of tagging is missing, thus making it difficult to design expert search approaches based on H5.

The appearance of this problem leads to a conclusion that some kind of guidelines and principles of good practice are obviously needed to guide the LOD set provides to avoid committing the above-mentioned types of errors, thus reduceing the usability of their data. We also believe that the recently emerged Pedantic Web[28] group might play a key role in making sure the data on the Web is given in a correct and useful form. One might also imagine the emergence of validators that would be able to verify not only the syntax of the given data, but also to check if the data fulfills the requirements of possible usage scenarios.

5.4 Lack of Interlinks

In some cases LOD is not a good source for expert finding because the datasets which may be used by certain hypothesis are not interlinked. During our evaluation we have found some examples of data that would be a useful source for expertise evidence if they would be interlinked.

Links to general topics are lacking for the majority of H1-related datasets, i.e. those that expose data about publications; as well as the H8-related datasets about research projects. Thus one may be able to find good experts who participated in funded research projects, but would not be able to correlate the projects with the appropriate generally used terms that identify topics.

Another example is the SW Conference data Corpus where a links exist to FOAF profiles, allowing one to relate a person with the papers he/she has published, and with professional events that the person has attended (as evidences of his/her competence). It is further possible to find the domains of expertise related to the

[17] http://oauth.net/

[18] http://www.data.gov/

[19] http://www.data.gov.uk/

[20] http://blogs.talis.com/nodalities/category/podcast

[21] http://videolectures.net

[22] http://www.slideshare.net

[23] http://swaml.berlios.de/

[24] http://www.twitter.com

[25] http://facebook.com

[26] http://expertfinder.info/

[27] http://www.faviki.com/

[28] http://pedantic-web.org/

research papers (thanks to the link with DBPedia concepts), but it is not possible to do the same for the professional events due to the lack of links to general topics.

Another important example are SIOC sites that would represent an excellent source of data for H17, thanks to SIOC exporters for Wordpress[29]. However, the tags that help identify the topic of the blog content are not always present. Fortunately Search Engine Optimization can be a good motivation for content producers to tag their blogs or use some automatic semantic tagging tools (e.g. Zemanta[30] and OpenCalais[31]).

DoapStore[32] is a promising source for H6-based approaches. It contains data on software development projects and their participants. Although the programming language data are present, they are only given in form of literals, and the presence of links to some general concepts (e.g. DBPedia or Freebase[33] ones) is not common. However, as we already stated, the H6-based approaches may rely on RDFOhloh for a more complete support. RDFOhloh also provides direct links to DoapStore descriptions, thus making the integration possible despite the lack of links in DoapStore.

Data needed for H8 is present in the Cordis dataset that is about all the European projects and the researchers involved. However, the data set uses its own representation of topics, and does not link to any general categories data sets (DBPedia, FreeBase, etc.).

As we have emphasized in the examples, the major obstacle for higher degree of linking among LOD datasets is related to the identity resolution problem – how to identify that two resources (either human or digital) are the same. This research challenge is known as "equivalence mining" within the LOD research community and there has already been significant amount of work directed at resolving it[34]. Reliance on inverse functional properties (e.g., foaf:mbox, foaf:homepage) is the most common approach. However, these properties are not always present in the description of resources. In such cases, establishing links between data is based on comparing labels (e.g., foaf:name, dc:title, etc) using different probabilistic and statistical methods (as shown in, e.g. [21]). Even more difficult problem that the research community has started to tackle is when data is represented using different, but comparable ontologies [22]. The Silk Framework[35] defines declarative language for specifying conditions that data items must fulfill in order to be interlinked and thus can be used in situations where terms from different ontologies are used and where no consistent RDFS or OWL scheme exists. It is our expectation that the intensifying Linked Data community effort in this area will result in highly interlinked LOD cloud that is highly conducible for expert search. It is also interesting to observe the emergence of new services such as Uberblic[36] that allow users to create their own equivalence mapping and thus infuse the new links in their view of the LOD cloud. This gives hope that links, which are the most important asset of LOD cloud, might be crowdsourced.

6. CONCLUSION

Expert search and profiling systems aggregate and analyze certain types of data depending on the types of expertise hypotheses they use. Traditional approaches tend to retrieve their data from closed or limited data corpuses. LOD on the other hand allows querying the whole Web like a huge database, thus surpassing the limits of closed data sets, and closed online communities. We believe that this opens new possibilities for traditional expert search and profiling systems which usually only rely on data from their local and limited databases or on unstructured data gathered from the Web. LOD also stands up for a great promise to deliver mutli-purpose data that can be used to find experts in many domains and with many different expertise hypotheses. In this paper we have explored the potentials and drawbacks of LOD in comparison to traditional data sources used for expert search,. We haven't only asked the question what LOD can do for you, but also what you can do for LOD to make it an even better source of expertise evidence. In general, the publishers of Linked Open Data should at least make sure:

- To publish the relevant evidence of expertise, with all the details that may be useful for finding and ranking experts;
- To provide a way to correlate a certain user (the expert candidate) with the evidence of competence and uniquely identify the user in other data sources (e.g. using inverse functional properties and owl:sameAs links);
- To provide a way to merge the data about an evidence of expertise from various data sources. For example one should be able to identify the same research paper in different data sources;
- To provide a way to correlate an evidence of expertise with recognizable and generally used terms that identify domains of expertise (e.g. DBPedia or Freebase concepts); and
- To provide means of authorization and protected access to privacy-sensible Linked Data.

Given the existing benefits of the current LOD cloud for experts search, as well as the easily attainable possibilities for its improvement, we remain strongly convinced in the bright future of LOD-based expert search approaches, that would be able to capture the essence of human knowledge, experience and activities through the traces that they leave on the Web, and evaluate their expert capabilities in a way that was not possible in the age before Linked Data.

In our future work, we will continue to develop a conceptual framework for expert search using LOD, and will develop services that could provide lists of experts, by running various hypotheses over LOD. The services will rely on our mapping of data sources and evidence types, and will employ the most recent tools for navigating through LOD developed by the Linked Data research community. We will also examine how expert finding can be coupled with problem solving communities, like Hypios.com.

---

[29] http://sioc-project.org/wordpress/

[30] http://www.zemanta.com/

[31] http://www.opencalais.com/

[32] http://doapstore.org

[33] http://www.freebase.com/

[34] http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining

[35] http://www4.wiwiss.fu-berlin.de/bizer/silk/

[36] http://uberblic.org

7. ACKNOWLEDGMENTS

The work of Milan Stankovic has been partially funded by ANRT – the French National Association for Research and Technology; under the grant number CIFRE N 789/2009.

8. REFERENCES

[1] Sonnentag, S. (2000) Expertise at work: Experience and excellent performance, *International Review of Industrial and Organizational Psychology,* vol. 15, pp. 223-264

[2] McDonald, D. W. and Ackerman, S. M. (1998). Just Talk to Me: A Field Study of Expertise Location, In *Proceedings of CSCW `98*, Seattle, WA, pp. 315-324.

[3] Balog, K., & Rijke, M. d. (2006). Finding experts and their details in e-mail corpora. *International World Wide Web Conference*. Retrieved from http://portal.acm.org/citation.cfm?id=1136002.

[4] Buitelaar, P., & Eigner, T. (2008). Topic Extraction from Scientific Literature for Competency Management. *The 7th International Semantic Web Conference*. Karlsruhe, Germany.

[5] Demartini, G., & Niederée, C. (2008). Finding experts on the semantic desktop. *The 7th International Semantic Web Conference*. Karlsruhe, Germany. Retrieved from http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-403/ISWC_PICKME08.pdf#page=23.

[6] Demartini, G. (2007). Finding experts using wikipedia. In *Proceedings of the Workshop on Finding Experts on the Web with Semantics (FEWS2007) at ISWC/ASWC2007,* (pp. 33-41). Busan, South Korea, Retrieved from http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-290/paper03.pdf.

[7] Kolari, P., Finin, T., Lyons, K., & Yesha, Y. (2008). Expert Search using Internal Corporate Blogs. In *Workshop on Future Challenges in Expertise Retrieval, SIGIR 2008* (pp. 2-5). Retrieved from http://ilps.science.uva.nl/fCHER/files/fCHER_proceedings.pdf#page=9.

[8] Chua, S. J. (2007). Using web 2.0 to locate expertise. *IBM Centre for Advanced Studies Conference*. Retrieved from http://portal.acm.org/citation.cfm?id=1321250.

[9] Amitay, E., Carmel, D., Golbandi, N., Har'El, N., Ofek-Koifman, S., Yogev, S., et al. (2008). Finding people and documents, using web 2.0 data. In *Proceedings of the SIGIR 2008 Workshop on Future Challenges in Expertise Retrieval* (pp. 1-6). Retrieved from http://ilps.science.uva.nl/fCHER/files/slides/fcher.harel.pdf.

[10] Adamic, L., Zhang, J., Bakshy, E., & Ackerman, M. (2008). Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web* (pp. 665-674). Beijing, China: ACM New York, NY, USA. Retrieved from http://portal.acm.org/citation.cfm?id=1367497.1367587.

[11] Jurczyk, P., & Agichtein, E. (2007). Discovering authorities in question answer communities by using link analysis. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*. New York, New York, USA: ACM Press. doi: 10.1145/1321440.1321575.

[12] Zhang, J., Ackerman, M. S., & Adamic, L. (2007). Expertise Networks in Online Communities: Structure and Algorithms. In *USA, 2004. ACM Press. WWW '07: Proceedings of the 16th international conference on World Wide Web* (pp. 221-230). New York, NY, USA: ACM Press.

[13] Noll, M. G., Yeung, C. A., Gibbins, N., Meinel, C., & Shadbolt, N. (2009). Telling Experts from Spammers: Expertise Ranking in Folksonomies. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. Boston, MA: ACM, New York, USA.

[14] Becerra-Fernandez, I. "Searching for experts on the web: a review of contemporary expertise locator systems," ACM Trans. on Internet Technology, vol. 6, no. 4, pp. 333-355, Nov. 2006.

[15] Fu, Y., Xiang, R., Liu, Y., & Zhang, M. (2007). Finding Experts Using Social Network Analysis. *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*. IEEE. doi: 10.1109/WI.2007.14.

[16] Yimam, D. (2000). Expert Finding Systems for Organizations: Domain Analysis and the DEMOIR Approach. ECSCW 99 Beyond Knowledge Management: Management Expertise Workshop. pp. 276-283

[17] Story, H., Harbulot, B., Jacobi, I., & Jones, M. (2009). FOAF+SSL: RESTful Authentication for the Social Web. In *Proceedings of SPOT2009, 1st Workshop on Trust and Privacy on the Social and Semantic Web* (p. Heraklion, Grece). Retrieved from http://ceur-ws.org/Vol-447/paper5.pdf.

[18] Brinker, S. 2010. 7 business models for linked data. [Online]. Available at: http://www.chiefmartec.com/2010/01/7-business-models-for-linked-data.html

[19] Dodds, L. 2010. Thoughts on Linked Data Business Models. [Online]. Available at: http://www.ldodds.com/blog/2010/01/thoughts-on-linked-data-business-models/

[20] Pellegrini, T. 2009. Linked Data Flows: A new picture to illustrate the "openness" we mean. [Online]. Available at: http://blog.semantic-web.at/2009/10/28/linked-data-flows-a-new-picture-to-illustrate-the-openness-we-mean/

[21] Tang, J., Zhang, J., Zhang, D., Yao, L., Zhu, C., Li, J., et al. (2007). ArnetMiner: An Expertise Oriented Search System for Web Community. In *Proceedings of Semantic Web Challenge'2007*.

[22] Nikolov, A., Uren, V., Motta. E. 2009. Towards Data Fusion in a Multi-ontology Environment. Proceedings of the WWW2009 Workshop on Linked Data on the Web, Madrid, Spain. Avaialble at: http://ceur-ws.org/Vol-538/ldow2009_paper15.pdf