How the Semantic Web is Being Used: An Analysis of FOAF Documents

Li Ding*, Lina Zhou†, Tim Finin*, Anupam Joshi*
* Department of Computer Science and Electrical Engineering
† Department of Information Systems
University of Maryland, Baltimore County
Baltimore MD USA
{dingli1,zhoul,finin,joshi}@umbc.edu

***Abstract*— Semantic Web researchers have initially focused on the representation, development and use of ontologies but paid less attention to the social and structural relationships involved. The past year has seen a dramatic increase in the amount of published RDF documents using the Friend of a Friend (FOAF) vocabulary, providing a valuable resource for investigating how early Semantic Web adopters use this technology as well as build social networks. We describe an approach to identify, discover, and analyze FOAF documents. Over 1.5 million of FOAF documents are collected to show the variety and scalability of the web of FOAF documents. We analyzed the empirical usage of namespace and properties in the FOAF community, which helps the FOAF project in standardizing vocabularies. We also analyzed the social networks induced by those FOAF documents and revealed interesting patterns which can become powerful resource for outsourcing and justification of scientific knowledge.**

I. INTRODUCTION

The Semantic Web offers a promising solution to publishing information and services on the World Wide Web augmented with descriptions in a form that is easier for machines to process and understand. This will help Web agents to performing a variety of tasks on behalf of their users, such as *information discovery and integration* and *service negotiation and composition*. Information published in the Semantic Web languages (RDF and OWL) uses terms denoting classes and properties drawn from one or more *ontologies*. These ontologies are online RDF documents that declare a set of terms with unique URIs and further define them by asserting logical relationships and constraints among them.

Among a large number of ontologies that have been published on the Web, however, only a few are well populated, i.e., have any significant use. [1] Our recent investigation of the namespaces of well populated ontologies (see Table I) revealed that, besides the meta-level ontologies (i.e. RDF, RDFS, DAML and OWL), one of the best populated ontology is FOAF (Friend-of-a-Friend) [1]. In addition, representing personal information is also a popular theme in ontology engineering (more than 1,000 RDF documents has defined terms containing ‘person’ [2]). The other well populated non-meta ontologies in Table I include: DC (Dublin Core Element Set) [3], which defines document metadata properties without domain/range qualification, and RSS (RDF Site Summary), which is “a lightweight multipurpose extensible metadata description and syndication format” for annotating websites [4] [3].

TABLE I
EIGHT BEST POPULATED ONTOLOGIES (GENERATED IN JUNE,2004)

Onto. Name	**Namespace URI**	**# of Docs. Populated**
RDF	http://www.w3.org/1999/02/22-rdf-syntax-ns#	$>1,129,749$
FOAF	http://www.foaf-project.org/	$>1,126,002$
DC	http://purl.org/dc/elements/1.1/	$>1,117,433$
RDFS	http://www.w3.org/2000/01/rdf-schema#	$>1,129,749$
MCVB	http://webns.net/mvcb/	$>8,838$
RSS	http://purl.org/rss/1.0/	$>7,560$
vCard	http://www.w3.org/2001/vcard-rdf/3.0#	$>6,229$
Bio	http://purl.org/vocab/bio/0.1/	$>6,183$

FOAF provides an RDF/XML vocabulary to describe personal information [5], including name, mailbox, homepage URL, friends, and so on. FOAF documents then induces the “web of acquaintances” [6] and thus an implicit trust network to support such applications as knowledge outsourcing [7] and online communities [8].

The advances in FOAF vocabulary and applications highlight several challenging issues that must be addressed. For example, how can one assemble a collection of FOAF documents to support Semantic Web research? What are the common patterns of connections among FOAF documents? What terms in FOAF vocabulary are the most frequently used? What is the potential of FOAF in enabling and enhancing the intelligence of Web-based information systems? The current FOAF literature ([9], [5], [8], [10], [11], [6]) provides a vision and various models of how FOAF documents might be used to support Web-based information system under the assumption that FOAF documents are widely available. There is still a

Partial research support was provided by DARPA contract F30602-00-0591 and NSF awards ITR-IIS-0326460 and ITR-IIS-0325464. We greatfully acknowledge many contributions from and interactions with colleagues in the UMBC ebiquity research group.

[1]When a term in an ontology is used (e.g., an instance of a class is created, or a property used to assert a relationship), we say that the term (and its ontology) is being populated. This is similar to the use of *populate* to refer to adding actual data to a database.

[2]This is reported by our Swoogle (http://swoogle.umbc.edu), a RDF crawling and indexing engine [2].

[3]The use of RSS is increasing dramatically as of this writing and Swoogle has discovered approximate 80,000 RSS documents by September, 2004.

lack of an empirical investigation on the characteristics and structure of the growing body of millions of FOAF documents. This paper presents the first empirical result to answer the above questions based on a large collection (over 1.5 million) of real world FOAF documents harvested from the Web.

Our research on online FOAF documents consists of four steps: identification of FOAF documents, discovery of FOAF documents using software agents, extraction of person information, and fusion of person information based on the semantics of FOAF vocabulary. Using the statistics over this collection of FOAF documents, we describe the common properties and namespaces shared by the FOAF community. We hope that this analysis might help FOAF developers design and build better tools as well as inform novice FOAF users on how to create effective FOAF documents. Analyses of the social networks encoded in FOAF documents provides insight into some interesting structural patterns of the Semantic Web from the person perspective. The richness of profiles in FOAF documents allows us to further characterize social ties and identify friendship types.

A direct result of this study will be a friendship directory in the Semantic Web. Based on this directory, reputation and recommendation systems can be maintained to help people choose trustworthy information (or service) providers and propagate trust through friendship relations. For example, an agent helping a customer find a good but inexpensive restaurant in a region might sort the recommendations based on the distance to the recommender in the social network.

Friendship networks connected by FOAF relationships can provide insights into features and patterns of social networks in the Semantic Web and advance the theories and models of social structures. Friendship networks in the physical world have been long studied in the social science. A well known example is Milgram's small-world phenomenon [12] – the observation that everyone in the world can be reached through a short chain of social acquaintances. The concept gives rise to the famous phrase *six degrees of separation*, which has recently been applied to social network analysis in both physical and virtual environments (e.g., [13], [9]). Social relationship have been derived from the contextual information or domain knowledge (e.g. co-citation relationship [14]) indirectly using data mining techniques. In addition to social networks, the collection of FOAF documents can serve as valuable resource for Semantic Web research in the development and testing of trust models as well as trust propagation models [15].

As the first study along this line, this paper reflects the state of FOAF usage and identifies any potential problems to guide the future practice. It further contributes to the stabilization of individual terms in FOAF vocabulary. Using people as the bridge, FOAF can potentially link most of other kinds of things we describe in the Web, including documents they co-authored, research interest they shared, photos they shot together, and so on. Based on relationships represented in FOAF, we can identify online communities in a research area and even discover existing communities and the emergence of new communities. As the Semantic Web evolves, there will be opportunities to study social dynamics and apply the findings in this study to support Semantic Web applications.

The remainder of this paper is organized as follows. Section two presents a review of the literature concerning FOAF vocabulary and social network analysis. Section three introduces a novel approach to building FOAF documents collection and analyzing the structure of friendship networks in the Semantic Web. Section four uses descriptive statistics and social network analysis to present findings on components of FOAF documents and structural relationships among person profiles. Section five concludes with a discussion the findings of this study and their implications to the Semantic Web research and practice.

II. BACKGROUND

A. FOAF Document

A *FOAF document* publishes "Web homepages for people, groups, companies and other kinds of thing", and it is "written in XML syntax, and adopts the conventions of the Resource Description Framework (RDF)" [16]. The FOAF project [1] was initiated by Dan Brickley and Libby Miller. It enriches the expression of personal information and relationships. So it is a useful building block for creating information systems that support online communities [5].

The most important component of a FOAF document is the *FOAF vocabulary*, which is identified by the namespace URI *http://xmlns.com/foaf/0.1/*. The FOAF vocabulary defines both classes (e.g., *foaf:Agent*, *foaf:Person*, and *foaf:Document*) and properties (e.g., *foaf:name*, *foaf:knows*, *foaf:interests*, and *foaf:mbox*) grounded in RDF semantics. In contrast to a fixed standard, the FOAF vocabulary is managed in an *open source* manner, i.e., it is not stable and is open for extension [1] [4]. Therefore, inconsistent FOAF vocabulary usage is expected across different FOAF documents. Currently, a large amount of FOAF documents are contributed by the fast-growing 'blog' websites.

The practical significance of FOAF to information creators and consumers can be illustrated with a variety of applications [5], [10], which are summarized as follows:

To creators, FOAF is useful by

- Managing communities by offering a basic expression for community membership. Many communities have proliferated on the Web, ranging from companies through professional organizations to social groups.
- Expressing identity by allowing unique user IDs across applications and services without compromising privacy. For example, the *foaf:mbox_sha1sum* property contains the ASCII-encoded SHA1 hash of a mailbox URI (e.g., *mailto:finin@Umbc.edu*. In order to prevent others from faking an identity, the encoding is designed as a one-way mapping and cannot be trivially reverse-engineered to reveal the original email address.
- Indicating authorship. FOAF tools use digital signatures to associate an email address with a document. Specifically, OpenPGP is used, along with the new namespace *http://xmlns.com/wot/0.1/* to denote concepts forming a "web of trust". This associates a signature with the

[4] The lastest FOAF specification only lists one stable term –'homepage' and leaves many others in 'testing' or 'unstable' stages.

document itself and then specifies a signature for the linked document as part of a *rdfs:seeAlso* link. Thus, authorship information can be expressed both inside and outside of the concerned documents.

FOAF supports consumers by:

- Allowing provenance tracking and accountability [10]. On the Web, the sorce of information is just as important as the information itself in judging its crediblility. Provenance tracking RDF tools can tell where and when a piece of information is obtained. A practice common to the FOAF community is to attach the source URI to each RDF statement.
- Providing assistance to new entrants in a community. For example, people unfamiliar with a community can learn the structure and authority of a research area from the community's FOAF files.
- Locating people with common interests. Users tend to be have interests and values similar to those they desire in others [9]. Peer-to-peer relationships are an essential ingredient to collaboration, which is the driving force of online communities.
- Augmenting email filtering by prioritizing mail from trustable colleagues. Using the degree of trust derived from FOAF files, people can prioritize incoming email and thus filter out those with low trust values.

B. Social Networks on the Web

A social network consists of people (or organizations or other social entities) connected by a set of social relationships, such as friendship, co-working or information exchange [17]. Determining the properties online communities is the most straightforward application of social network analysis techniques. Microsoft NETSCAN maintains large online communities covering different topics and areas. The underlying physical social network can be reflected in an online community. For example, Club Nexus [9] is an online community serving over 2000 Stanford undergraduate and graduate students. Students can use Club Nexus to send email and invitations to events, post events, buy and sell goods, search and connect to people with similar interests, etc. The statistical analyses revealed that personalities and preferences of users mostly align with each other.

In addition to member relationship in online communities, social network analysis has been applied to many other types of social networks. For example, Xu and Chen [13] created, analyzed and visualized a network or criminals. Using social network analysis, they constructed patterns that represent criminal networks and associations between criminals automatically. The analysis not only yields the main groups but also identifies the subgroups, the key individuals (centrality) and links between groups. Centrality can be detected using degree (the number of direct links), betweenness (geodesics passing through), and closeness (sum of geodesics). Each of these indices infers different individual roles, a high degree infers leadership and betweenness infers a gatekeeper. This increased understanding enables officers to target specific criminals, to disrupt criminal organizations, and to achieve higher rates of conviction through associations.

Chen [14] describes the development and application of visualization techniques allowing users to access and explore information in a digital library effectively and intuitively based on co-citation relationships. Salient semantic structures and citation patterns are extracted from several document collections using latent semantic indexing and pathfinder network scaling. Author co-citation patterns are visualized through a number of author co-citation maps highlighting important research areas in the field. This approach provides a means of transcending the boundaries of collections of documents and visualizing more profound patterns in terms of semantic structures and co-citation networks.

Link structure analyses and graph-theory have been applied to crawling the Web for cyber-communities [18]. The FOAF project takes the social networking aspect of the Web further [8]. An interesting and powerful use of information in FOAF documents arises when data are aggregated and then cross-linked.

III. Building FOAF document collection

In order to build a collection of FOAF documents, we developed a multi-step process involving identification, discovery, and extraction. In the following context, we use the following *qualified names* [5] or QNames as short form (see also Table II).

TABLE II

QNames and corresponding namespaces

rdfs:	http://www.w3.org/2000/01/rdf-schema#
foaf:	http://xmlns.com/foaf/0.1/
dc:	http://purl.org/dc/elements/1.1/
bio:	http://purl.org/vocab/bio/0.1/
daml:	http://www.daml.org/2001/03/daml+oil#

A. FOAF Document Identification

It is easy to test if a RDF document uses the FOAF namespace, but it is non-trivial to give crisp criteria to determine whether a document is a FOAF document. A FOAF document does not have a fixed structure since it maybe published by different authors with various intensions. Typical FOAF documents are personal profiles on blog websites and collective person directory. Although FOAF vocabulary may be used to describe things than people, we focus on its usage in describing personal information.

In order to give a formal definition of FOAF document, we analyze the characteristic patterns which are implied by the ontological semantics and the empirical usage of FOAF vocabulary. Therefore, we define a **strict FOAF document** D with the following four characteristic patterns:

1) D is a valid RDF document. This can be validated by a RDF parser.

[5]In XML documents, a QNname is typically of the form *prefix:localName* where *prefix* is a symbol defined as a namespace corresponding to a full URL in the header of the document.

2) D uses the FOAF namespace.
3) D contains an RDF graph pattern as show in Figure 1. In this figure, X and Z are two different instances of *rdfs:Resource* and Y is an instance of *rdf:Property* using FOAF namespace.
4) D defined only one instance of *foaf:Person* without referencing it as *object* in any triples within D. D may additionally has some other instances of *foaf:Person*; however, each of them must be referenced as an *object* in at least one triple in D.

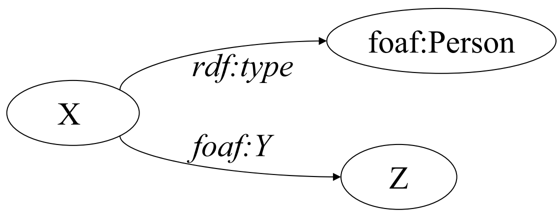


Fig. 1. FOAF document pattern

The above patterns, especially the fourth pattern, are quite strict and exclude many documents not dedicating to a person. Therefore, by removing the fourth pattern, we define a **general FOAF document** as long as it contains instances of *foaf:Person*.

Since RDF/XML is commonly used in encoding RDF, we only considered RDF/XML encoded RDF files in this study[6]. We noticed that an instance of *foaf:Person* usually instantiates some FOAF properties when describing a person. Based on the semantics of *rdfs:domain*, we simplified the operation of the third patten as testing the existence of FOAF properties whose *rdfs:domain* is *foaf:Person*.

B. FOAF Document Discovery

Based on the above patterns, we used both conventional web search engines and semantic crawlers to discover online FOAF documents. The FOAF document discovery is an iterative process consisting of two steps: (i) using web search engines to discover potential URLs of FOAF documents; and (ii) running the crawler to validate and discover new links according to FOAF vocabulary semantics.

Conventional web search engines provide a good starting point to discover FOAF documents on the Web [19]. Since web search engines treat Web documents as free text, we translate the four FOAF document patterns into effective queries. According to the first pattern, we only query documents which use well-known suffices for RDF documents, including ".rdf", ".xrdf", ".owl" and so on. Even though such suffixes were neither sufficient nor necessary for classifying RDF documents, we did obtain a significant number of RDF documents with high precision. According to the second and third patterns, we generated several search keywords. Table III shows the effectiveness (recall) of different query strings on different search engines. It is notable that Google outperformed other search engines due to its capability of searching file types. A practical issue with using web search engines is that they do not fully support meta search, i.e., they only return a small portion of Web documents that have been discovered. For instance, Google only returns the URLs of the first 1,000 documents for any query.

TABLE III

RETRUNED RESULTS FROM WEB SEARCH ENGINES

	google	yahoo	AskJeeves
filetype:rdf foaf	13,600	5,730	26
xmlns.com/foaf/0.1/	9,390	1	1,000+
foaf:Person	5,080	2,820	1,000+

The semantics of FOAF ontology indicates that FOAF documents are linked by *foaf:knows* augmented by *rdfs:seeAlso*. It is suggested in the FOAF specification that "Perhaps the most important use of *foaf:knows* is, alongside the *rdfs:seeAlso* property, to connect FOAF files together." This idea was reflected in our semantic crawler.

We initially discovered over 10,000 FOAF documents from conventional Web search engines. As the indexed data in search engines evolves, we may discover more over time. Using our FOAF crawler, which takes advantages of the semantics of FOAF vocabulary, we then discovered over 1.5 million FOAF documents[7]. Figure 2 shows how those FOAF documents are located across over 5,400 different websites. Table IV shows that most FOAF documents are contributed by several websites. Moreover, six of the seven top websites (except *www.ilrt.bris.ac.uk*) are *blog* websites[8], where FOAF documents are automatically generated.

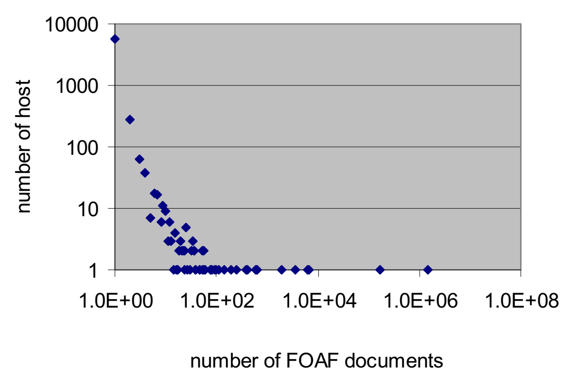


Fig. 2. FOAF documents distribution

C. Findings on Properties of Person

The FOAF project has tried to standardize the FOAF ontology through "endless" discussion. We believe that the standardization should take users' adoption into consideration.

[6]Ignoring, for example, FOAF documents encoded in N3.

[7]There are some commercial websites like http://linkedIn.com/ that use the FOAF ontology, but they protect their FOAF documents from the public access.

[8]According to WordNet2, *http://www.cogsci.princeton.edu/cgi-bin/webwn*, a *blog* is equivalent to a *web log*, which is "a shared on-line journal where people can post diary entries about their personal experiences and hobbies".

TABLE IV

TOP 7 HOST OF FOAF DOCUMENTS

host	disc. by us	disc. by Google
www.livejournal.com	1,508,346	0
www.deadjournal.com	169,546	0
www.ecademy.com	6,208	0
www.meinbild.ch	3,697	0
blog.livedoor.jp	1,914	3,560
www.ilrt.bris.ac.uk	621	651
eikeon.com	403	408

Therefore, in order to examine the properties of *foaf:Person*, we classify the colleted FOAF documents into two groups:

1) FOAF documents from blog websites. These are generated automatically from information entered via a online form or stored in a database. Documents from one host or website use the same vocabulary and syntactic structures. Such FOAF documents account for over 1.5 million files in our collection.
2) FOAF documents from non-blog websites. These are normally written by humans, and use a variety of vocabularies and structures. There are 5,000 documents from non-blog web sites in our collection.

The analyses of the general usage of properties in FOAF documents revealed the following findings(see Table V): (i) people have privacy concerns when they supply their personal information. For example, *foaf:mbox_sha1sum* and *foaf:homepage* are considered less private than *foaf:mbox* and *foaf:dateOfBirth*, and thus the former were used more frequently than the latter. (ii) Properties of a person may come from ontologies other than FOAF such as 'bio' and 'dc'. (iii) FOAF documents are fairly connected because of the heavy usage of the combination of *foaf:knows* and *rdfs:seeAlso*. We also observed that FOAF documents from livejournal.com is better connected than those from non-blog websites.

TABLE V

10 MOST USED PROPERTIES (PER FOAF FILE)

from non-blog (5232)	from livejournal.com (169,115)
foaf:mbox_sha1sum (0.84)	foaf:mbox_sha1sum (1.0)
foaf:homepage (0.66)	dc:description(1.0)
foaf:name (0.64)	dc:title (1.0)
foaf:nick (0.61)	foaf:nick (1.0)
foaf:weblog (0.60)	foaf:page (1.0)
foaf:knows (0.44)	foaf:weblog (0.99)
foaf:mbox (0.38)	rdfs:seeAlso (0.85)
foaf:img (0.38)	foaf:knows (0.85)
bio:olb (0.35)	foaf:dateOfBirth (0.71)
rdfs:seeAlso (0.34)	foaf:interest (0.67)

Based on the usage of properties of person in *foaf:Person* instances(see table VI), we observed the following patterns: (i) although both the *foaf:title* property and the *dc:title* property were used, *dc:title* was used more frequently; and (ii) The *foaf:nick* property was used much more frequently in blog websites than non-blog ones, for it served as a unique identifier of a person.

TABLE VI

10 MOST USED PROPERTIES(PER INDIVIDUAL/INSTANCE)

from non-blog (26,936)	from liveJournal.com (20,298,073)
foaf:name (0.84)	dc:title (1.74)
foaf:knows (0.79)	foaf:interest (1.68)
foaf:homepage (0.63)	foaf:nick (1.04)
foaf:mbox_sha1sum (0.51)	foaf:weblog (1.00)
rdfs:seeAlso (0.40)	rdfs:seeAlso (0.99)
dc:title (0.31)	foaf:knows (0.95)
foaf:nick (0.22)	foaf:page (0.95)
foaf:weblog (0.18)	dc:description (0.046)
foaf:mbox (0.15)	foaf:mbox_sha1sum (0.046)
daml:equivalentTo (0.13)	foaf:dateOfBirth (0.046)

IV. APPLICATION OF FOAF

A. *Retrieval and Fusion of Personal Information*

One of the principles of the semantic web is that "anyone is allowed to say anything about any resource". For example, document $D1$ can make assertions about individuals introduced in document $D2$. Since FOAF is based on RDF, this allows one person to assert information about others, be they friends, acquaintances or complete strangers. Therefore, we can retrieve information from a collection of FOAF documents about a person even if he has not published his own FOAF document. For example, a FOAF document[9] may provide information about the individual with name "Dr. Benjamin Grosof" even if though he has not published his FOAF document. Moreover, information about an individual may be spread across a number of FOAF documents in a collection, providing a kind of community view that mirrors the person's view in the community of people.

Property	Value
foaf:mbox	mailto:bgrosof@mit.edu
foaf:name	Benjamin Grosof
foaf:mbox_sha1sum	4a5be6ce326bfa04a7cb29a77e944795b93ce7b4

When a person is described in more than one FOAF documents, we must fuse information from multiple sources and generate aggregated information about the person. The FOAF ontology semantics defines unique identifiers of person, such as *foaf:mbox*, *foaf:mbox_sha1sum*, and *foaf:homepage*, which are ideal clues to information fusion. For example, Dr. Tim Finin's personal information can be obtained not only from his own FOAF document and but also from references in FOAF documents authored by other people (see Figure 3). We found two different values of *foaf:name* in this case. It gave us a clue of Tim Finin's name, "either *Tim Finin* or *Timothy W. Finin*", which is the unique identifier of an author in DBLP[10].

Caution should be taken in merging information from multiple FOAF documents since some of the facts may be wrong and the collection of facts may contain contradictions. Errors in FOAF documents can lead to unexpected results. For example, after fusing personal information of Dr. Jim Hendler with his email address *hendler@cs.umd.edu*,

[9] 'http://www.ilrt.bris.ac.uk/discovery/2001/06/content/rdf_meeting.rdf'
[10] http://www.informatik.uni-trier.de/ ley/db/

we found a mixture of information from two different persons. Figure 4 lists three names in the fusing result: "Jim Hendler","Norman Walsh", and "Norm Walsh". Our investigation revealed that the errors were caused by the document *http://www.ilrt.bris.ac.uk/people/cmdjb/webwho.xrdf*, in which *foaf:mbox_sha1sum* was mistakenly associated with Norman's email-hash[11].

B. Social Network Analysis

A collection of distributed FOAF documents may constitute a social network. The *foaf:knows* relation can link one individual of *foaf:Person* to another. By matching the unique identifiers (e.g. *foaf:mbox*) defined by FOAF vocabulary, we were able to obtain *owl:sameIndividualAs* relations among individuals of *foaf:Person* defined in different FOAF documents, and unify those RDF nodes to build a big social network graph. We focused on emerging social networks in the distributed Semantic Web, which is different than those emerged from a centralized blog website. Therefore, our analyses only concerned FOAF documents from non-blog websites. We found many instances followed Zipf's distributions [20], so all the figures of distribution were plotted on log-log scale.

1) Size of FOAF Social Networks: We selected about 7,000 FOAF documents containing 50,559 instances of *foaf:Person*. Figure 5 shows the distribution of individuals of *foaf:Person* defined in FOAF documents. The average number of persons mentioned by a FOAF document was 7.22, while the median was 1. This shows that more than half (56%) of the authors of FOAF documents provided information about themselves without including information about their social network (acquaintances or friends). These documents are strict FOAF documents.

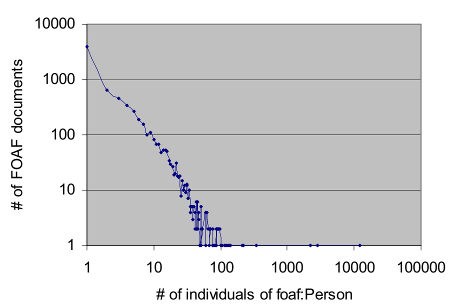


Fig. 5. Distribution of persons per FOAF document

Mapping between people is crucial to building a connected social network. Following to RDF and FOAF semantics, we consider several types of mappings:

- Multiple identities of a person are grouped by an instance of *foaf:Person*. For example, a FOAF document may list several values of *foaf:mbox* for the same person.
- Same identity across FOAF documents. For example, instances of *foaf:Person* sharing the same unique person identifier (e.g. *foaf:mbox_sha1sum*, *foaf:mbox*, and *foaf:homepage*) must refer to the same person.
- URL of strict FOAF document. Given an instance of *foaf:Person*, the values of *rdfs:seeAlso* properties are usually URLs of strict FOAF documents that extensively describe the same person.

In our preliminary study, we grouped people according to their *foaf:mbox_sha1sum*. As a result, we found 42,504 distinctive groups of *foaf:Person* individuals. Figure 6 shows that only 6.3% groups have more than two persons. People associated with big groups are, in fact, social authorities, who are known by many people. In the discussion there forth, we use "group" as a fused person.

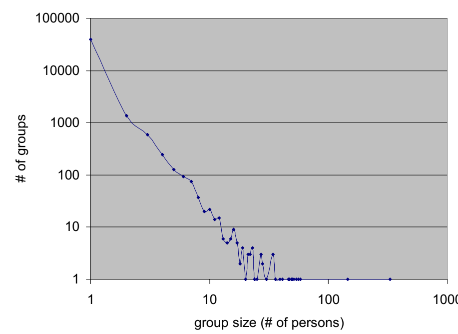


Fig. 6. Distribution of persons per group

2) Patterns of Degree: Degree analysis is an important tool in social network analysis. Our analyses were based on 35,299 'knows' links within non-blog documents. The blog web sites may introduce many more links (our crawler had discovered more than 20,371,514 'knows' relations). Figure 8 and 7 display the distributions of in-degrees and out-degrees respectively. It is shown that only a few groups had more than one in-degrees or out-degrees. Our statistics on in-degrees and out-degrees (See figure 9) also showed that, among the 42,504 groups, only 7% of them have both in-links and out-links. Moreover, 97.7% of the 'only in' nodes have only one in-link. These distributions of node degree revealed the sparseness of groups.

3) Patterns of Connected Components: The connected components generated from the FOAF documents exhibits interesting graphical patterns. We identified three basic types of patterns:

- **star**. The *foaf:knows* relation produces directed graphs, thus the star pattern is essentially produced by a active person who knows many other people.
- **clique**. The clique pattern emerges when there is a small group of friends who have FOAF documents that reference many of the other clique members.
- **singleton**. The singleton pattern corresponds to a person who has no links to other people. Such a pattern is likely to be generated by someone experimenting with the creation of a FOAF document.

[11]Norman's email-hash is *ef99fd659575b85b94575cc016043813ec1294dc* according to *http://norman.walsh.name/knows/who#norman-walsh*

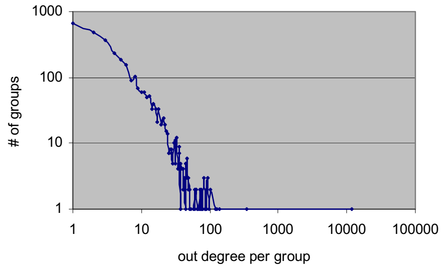


Fig. 7. out-degree distribution per group

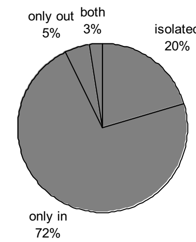


Fig. 9. Distribution of Patterns of node's degree

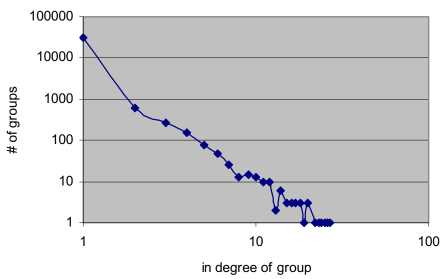


Fig. 8. In-degree distribution per group

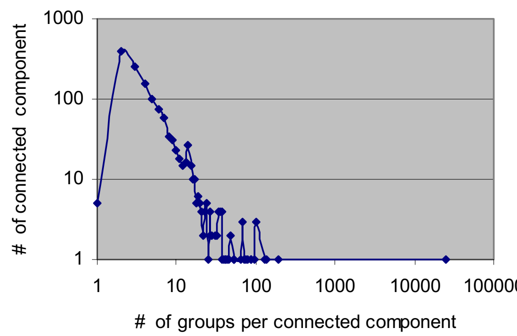


Fig. 10. Distribution of groups per connected component

Among the 1,337 connected components that we have discovered, most of them have fewer than five groups, as shown in Figure 10. We also found two types of extreme cases of connected components: (i) those singletons each of which contains only one group; and (ii) one that has 24,559 groups. This is produced by a document that has defined more than 10k instances of *foaf:Person*. Figure 11[12] visualizes how groups are connected through different connected components.

We hypothesis that the topology of FOAF network would evolve over time. As shown in Figures 12, a connected component (CC) formed into a star at the beginning (Fig. 12a), then it was connected to another star (Fig. 12b), and so forth. Finally, some cliques joined the connected component, and consequently the CC continued to grow bigger and bigger.

V. Conclusion

We presented a novel perspective of the Semantic Web by linking machine-readable descriptions of people, i.e. FOAF documents, with published personal relationships. This complements the ontology-based view of the Semantic Web. We also proposed a heuristic approach to identifying and discovering FOAF documents from the Web and extracting information about people from these FOAF documents. This approach provides a means of transcending the boundaries of individual FOAF documents, fusing information about a person from multiple documents. The analysis of FOAF network pattern also lent itself to unique social network structures in the Semantic Web.

FOAF networks provide a snapshot of the FOAF user community encoded in the constituent *foaf:knows* relations. More importantly, connection patterns among FOAF documents offer a persons orientation to the conventional Web of HTML documents. The visualization of highly connected FOAF networks is informative and revealing. As the number of FOAF users grows, the approach presented in this paper can be used to discover existing and emerging online communities.

Our investigation of personal information related ontologies clearly delineated two types of sources for FOAF documents: those generated from blog sites and those generated by other tools or manually. The statistics on FOAF documents from non-blog sources, rather than blog sources, best revealed the choices made by the Semantic Web community, including the most commonly used properties and namespaces. These findings provide guidance to stabilize FOAF vocabulary as well as authorize FOAF documents.

The collection of FOAF documents itself can become valuable resource for researchers to study the (social) network structure in the Semantic Web. More work is needed in analyzing FOAF documents and discovering patterns in connected components. We will expand the scope of links for building networks to include implicit relationships such as

[12] Figure 11 and 12 were generated by the "Otter" network visualization tool [21].

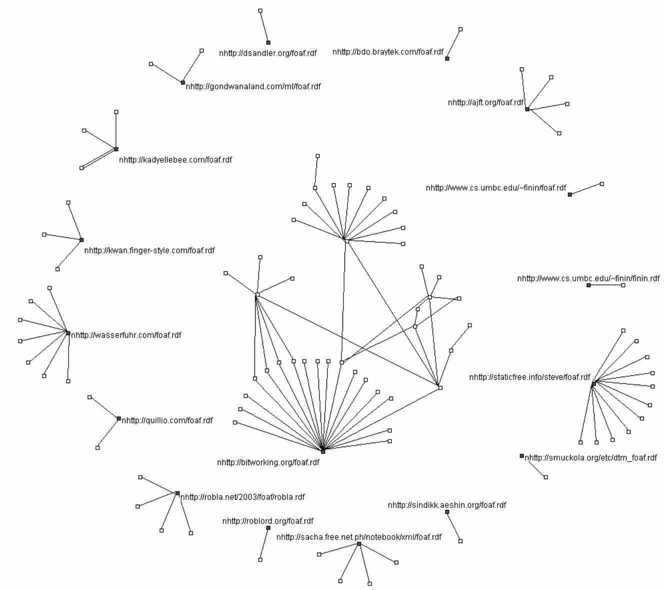
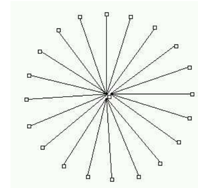
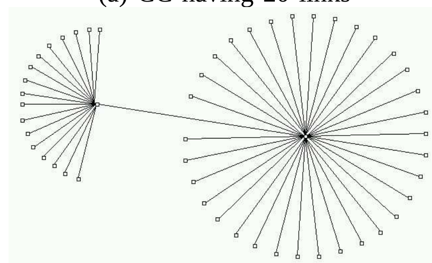


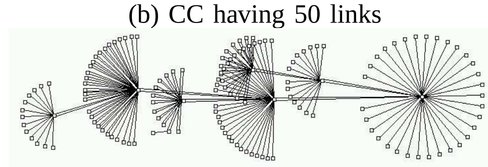
Fig. 11. Some connected components in FOAF network.



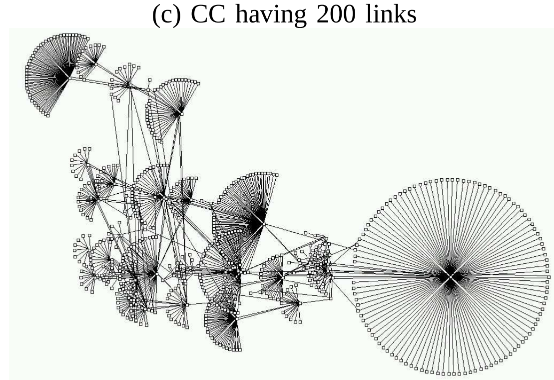
(a) CC having 20 links



(b) CC having 50 links



(c) CC having 200 links



(d) CC having 1000 links

Fig. 12. The evolution of a connected component.

common interest and experience. We will refine the approach to FOAF network analysis in order to provide more effective and informative representation of the structure of the Semantic Web. We plan to select a knowledge domain to analyze the structure FOAF network, which is expected to serve as a powerful means for outsourcing and justification of scientific knowledge.

REFERENCES

[1] "The friend of a friend(foaf) project," http://www.foaf-project.org/.

[2] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi, , and J. Sachs, "Swoogle: A search and metadata engine for the semantic web," in *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, 2004.

[3] "The dublin core element set v1.1 namespace providing access to its content by means of an rdf schema," http://purl.org/dc/elements/1.1/.

[4] "Rdf site summary," http://web.resource.org/rss/1.0/spec.

[5] E. Dumbill, "Finding friends with xml and rdf," IBM's XML Watch, http://www-106.ibm.com/developerworks/xml/library/x-foaf.html, June 2002.

[6] J. Golbeck, B. Parsia, and J. Hendler, "Trust networks on the semantic web," in *Proceedings of Cooperative Intelligent Agents*, 2003.

[7] L. Ding, L. Zhou, and T. Finin, "Trust based knowledge outsourcing for semantic web agents," in *Proceedings of IEEE/WIC International Conference on Web Intelligence*, 2003.

[8] E. Dumbill, "Support online communities with foaf: How the friend-of-a-friend vocabulary addresses issues of accountability and privacy," IBM's XML Watch, http://www-106.ibm.com/developerworks/xml/library/x-foaf2.html, August 2002.

[9] L. A. Adamic, O. Buyukkokten, and E. Adar, "A social network caught in the web," *First Monday*, vol. 8, no. 6, June 2003.

[10] E. Dumbill, "Tracking provenance of rdf data," IBM's XML Watch, http://www-106.ibm.com/developerworks/xml/library/x-rdfprov.html, July 2003.

[11] G. A. Grimnes, P. Edwards, and A. Preece, "Learning meta-descriptions of the foaf network," in *Proceedings of International Semantic Web Conference*, 2004.

[12] S. Milgram, "The small world problem," *Psychology Today*, vol. 1, no. 1, pp. 60–67, 1967.

[13] J. Xu and H. Chen, "Untangling criminal networks: A case study," in *Intelligence and Security Informatics, First NSF/NIJ Symposium*, vol. 2665, 2003, pp. 232–248.

[14] C. Chen, "visualising semantic spaces and author co-citation networks in digital libraries," *Inf. Process. Manage.*, vol. 35, no. 3, pp. 401–420, 1999.

[15] L. Ding, P. Kolari, S. Ganjugunte, T. Finin, , and A. Joshi, "Modeling and evaluating trust network inference," in *Seventh International Workshop on Trust in Agent Societies at AAMAS 2004*, 2004.

[16] "Foaf vocabulary specification version 1.0(foaf galway edition)," http://xmlns.com/foaf/0.1/.

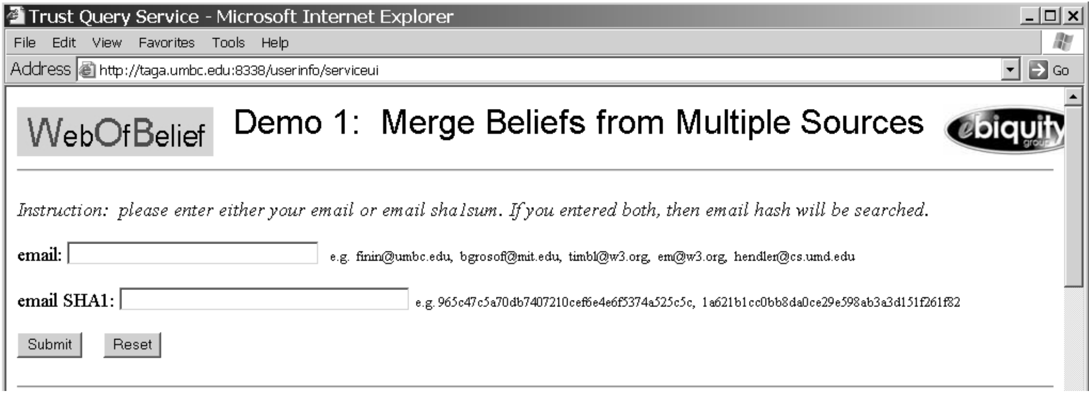
[17] L. Garton, C. Haythornthwaite, and C. Haythornthwaite, "Studying online social networks," *Journal of Computer-Mediated Communication*, vol. 3, 1997.

[18] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the web for emerging cyber-communities," in *Proceeding of the eighth international conference on World Wide Web*. Elsevier North-Holland, Inc., 1999, pp. 1481–1493.

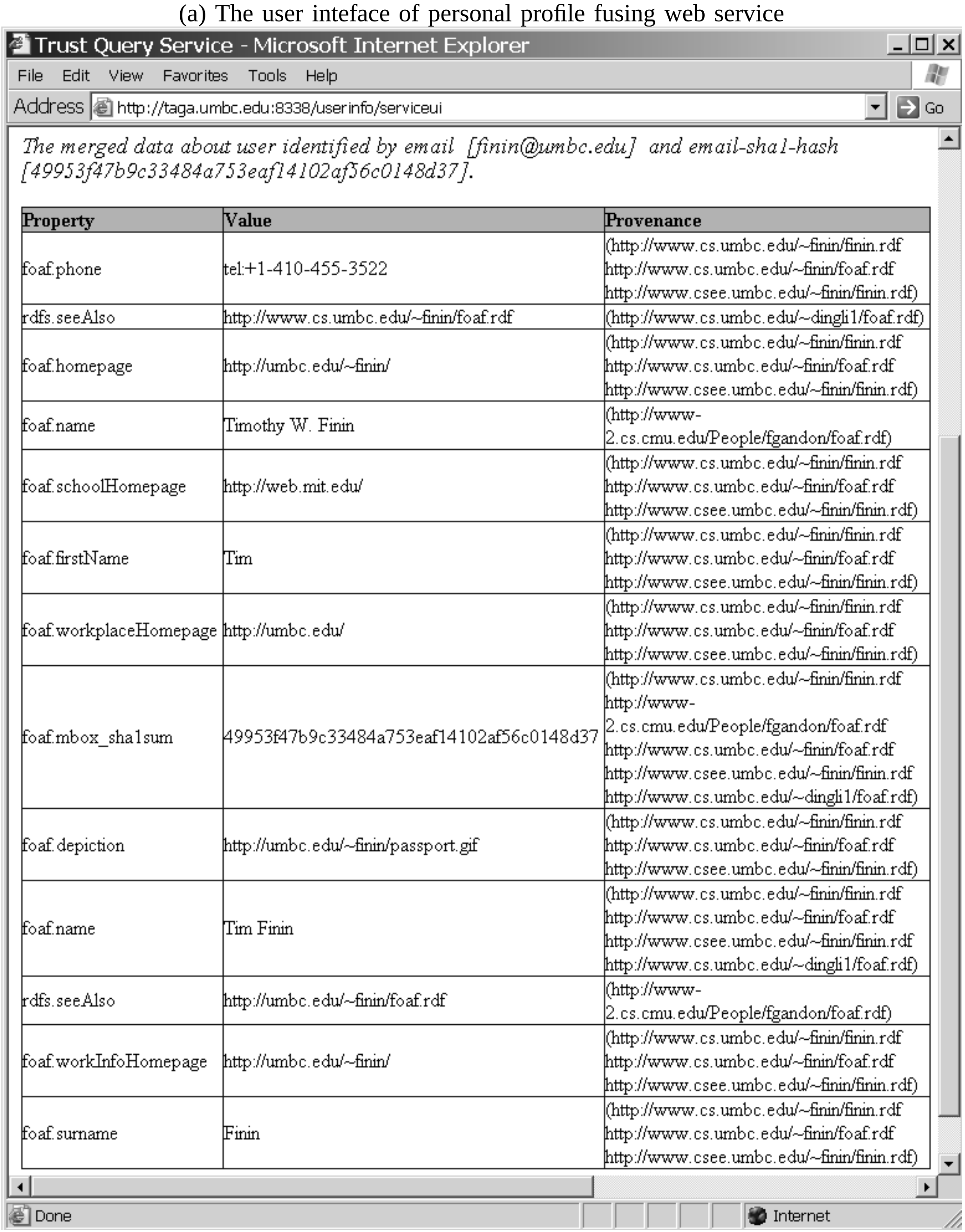
[19] B. DuCharme, "Googling for xml," http://www.xml.com/pub/a/2004/02/11/googlexml.html, 2004.

[20] G. K. Zipf, *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, 1932.

[21] B. Huffaker, "Otter: A general-purpose network visualization tool." [Online]. Available: http://www.caida.org/tools/visualization/otter/



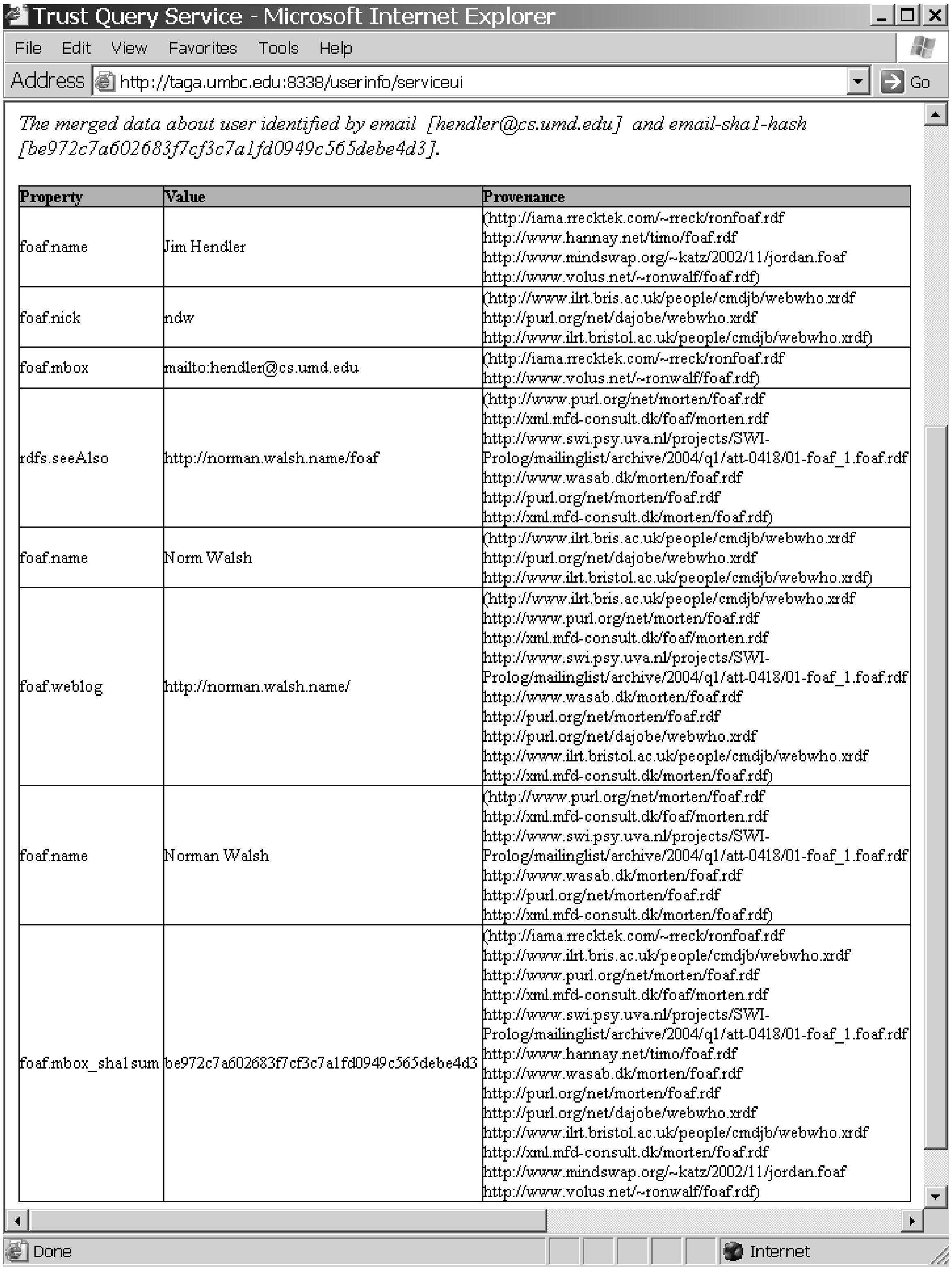
(a) The user inteface of personal profile fusing web service



Property	Value	Provenance
foaf.phone	tel:+1-410-455-3522	(http://www.cs.umbc.edu/~finin/finin.rdf http://www.cs.umbc.edu/~finin/foaf.rdf http://www.csee.umbc.edu/~finin/finin.rdf)
rdfs.seeAlso	http://www.cs.umbc.edu/~finin/foaf.rdf	(http://www.cs.umbc.edu/~dingli1/foaf.rdf)
foaf.homepage	http://umbc.edu/~finin/	(http://www.cs.umbc.edu/~finin/finin.rdf http://www.cs.umbc.edu/~finin/foaf.rdf http://www.csee.umbc.edu/~finin/finin.rdf)
foaf.name	Timothy W. Finin	(http://www-2.cs.cmu.edu/People/fgandon/foaf.rdf)
foaf.schoolHomepage	http://web.mit.edu/	(http://www.cs.umbc.edu/~finin/finin.rdf http://www.cs.umbc.edu/~finin/foaf.rdf http://www.csee.umbc.edu/~finin/finin.rdf)
foaf.firstName	Tim	(http://www.cs.umbc.edu/~finin/finin.rdf http://www.cs.umbc.edu/~finin/foaf.rdf http://www.csee.umbc.edu/~finin/finin.rdf)
foaf.workplaceHomepage	http://umbc.edu/	(http://www.cs.umbc.edu/~finin/finin.rdf http://www.cs.umbc.edu/~finin/foaf.rdf http://www.csee.umbc.edu/~finin/finin.rdf)
foaf.mbox_sha1sum	49953f47b9c33484a753eaf14102af56c0148d37	(http://www.cs.umbc.edu/~finin/finin.rdf http://www-2.cs.cmu.edu/People/fgandon/foaf.rdf http://www.cs.umbc.edu/~finin/foaf.rdf http://www.csee.umbc.edu/~finin/finin.rdf http://www.cs.umbc.edu/~dingli1/foaf.rdf)
foaf.depiction	http://umbc.edu/~finin/passport.gif	(http://www.cs.umbc.edu/~finin/finin.rdf http://www.cs.umbc.edu/~finin/foaf.rdf http://www.csee.umbc.edu/~finin/finin.rdf)
foaf.name	Tim Finin	(http://www.cs.umbc.edu/~finin/finin.rdf http://www.cs.umbc.edu/~finin/foaf.rdf http://www.csee.umbc.edu/~finin/finin.rdf http://www.cs.umbc.edu/~dingli1/foaf.rdf)
rdfs.seeAlso	http://umbc.edu/~finin/foaf.rdf	(http://www-2.cs.cmu.edu/People/fgandon/foaf.rdf)
foaf.workInfoHomepage	http://umbc.edu/~finin/	(http://www.cs.umbc.edu/~finin/finin.rdf http://www.cs.umbc.edu/~finin/foaf.rdf http://www.csee.umbc.edu/~finin/finin.rdf)
foaf.surname	Finin	(http://www.cs.umbc.edu/~finin/finin.rdf http://www.cs.umbc.edu/~finin/foaf.rdf http://www.csee.umbc.edu/~finin/finin.rdf)

(b) The fused personal profile with proveance information

Fig. 3. Fusing Dr. Tim Finin's person information



The merged data about user identified by email [hendler@cs.umd.edu] and email-sha1-hash [be972c7a602683f7cf3c7a1fd0949c565debe4d3].

Property	Value	Provenance
foaf.name	Jim Hendler	(http://iama.rrecktek.com/~rreck/ronfoaf.rdf http://www.hannay.net/timo/foaf.rdf http://www.mindswap.org/~katz/2002/11/jordan.foaf http://www.volus.net/~ronwalf/foaf.rdf)
foaf.nick	ndw	(http://www.ilrt.bris.ac.uk/people/cmdjb/webwho.xrdf http://purl.org/net/dajobe/webwho.xrdf http://www.ilrt.bristol.ac.uk/people/cmdjb/webwho.xrdf)
foaf.mbox	mailto:hendler@cs.umd.edu	(http://iama.rrecktek.com/~rreck/ronfoaf.rdf http://www.volus.net/~ronwalf/foaf.rdf)
rdfs.seeAlso	http://norman.walsh.name/foaf	(http://www.purl.org/net/morten/foaf.rdf http://xml.mfd-consult.dk/foaf/morten.rdf http://www.swi.psy.uva.nl/projects/SWI-Prolog/mailinglist/archive/2004/q1/att-0418/01-foaf_1.foaf.rdf http://www.wasab.dk/morten/foaf.rdf http://purl.org/net/morten/foaf.rdf http://xml.mfd-consult.dk/morten/foaf.rdf)
foaf.name	Norm Walsh	(http://www.ilrt.bris.ac.uk/people/cmdjb/webwho.xrdf http://purl.org/net/dajobe/webwho.xrdf http://www.ilrt.bristol.ac.uk/people/cmdjb/webwho.xrdf)
foaf.weblog	http://norman.walsh.name/	(http://www.ilrt.bris.ac.uk/people/cmdjb/webwho.xrdf http://www.purl.org/net/morten/foaf.rdf http://xml.mfd-consult.dk/foaf/morten.rdf http://www.swi.psy.uva.nl/projects/SWI-Prolog/mailinglist/archive/2004/q1/att-0418/01-foaf_1.foaf.rdf http://www.wasab.dk/morten/foaf.rdf http://purl.org/net/morten/foaf.rdf http://purl.org/net/dajobe/webwho.xrdf http://www.ilrt.bristol.ac.uk/people/cmdjb/webwho.xrdf http://xml.mfd-consult.dk/morten/foaf.rdf)
foaf.name	Norman Walsh	(http://www.purl.org/net/morten/foaf.rdf http://xml.mfd-consult.dk/foaf/morten.rdf http://www.swi.psy.uva.nl/projects/SWI-Prolog/mailinglist/archive/2004/q1/att-0418/01-foaf_1.foaf.rdf http://www.wasab.dk/morten/foaf.rdf http://purl.org/net/morten/foaf.rdf http://xml.mfd-consult.dk/morten/foaf.rdf)
foaf.mbox_sha1sum	be972c7a602683f7cf3c7a1fd0949c565debe4d3	(http://iama.rrecktek.com/~rreck/ronfoaf.rdf http://www.ilrt.bris.ac.uk/people/cmdjb/webwho.xrdf http://www.purl.org/net/morten/foaf.rdf http://xml.mfd-consult.dk/foaf/morten.rdf http://www.swi.psy.uva.nl/projects/SWI-Prolog/mailinglist/archive/2004/q1/att-0418/01-foaf_1.foaf.rdf http://www.hannay.net/timo/foaf.rdf http://www.wasab.dk/morten/foaf.rdf http://purl.org/net/morten/foaf.rdf http://purl.org/net/dajobe/webwho.xrdf http://www.ilrt.bristol.ac.uk/people/cmdjb/webwho.xrdf http://xml.mfd-consult.dk/morten/foaf.rdf http://www.mindswap.org/~katz/2002/11/jordan.foaf http://www.volus.net/~ronwalf/foaf.rdf)

Fig. 4. Fusing Dr. Jim Hendler's person information