Finding People and Organizations on the Semantic Web

Jussi Kurki*
* Semantic Computing Research Group (SeCo)
Helsinki University of Technology (TKK) and University of Helsinki
first.last@tkk.fi

Abstract

Finding people is essential in finding information. Librarians and information scientists have studied authority control - psychologists and sociologists social networks. In aforementioned, authors link to documents (and co-authors) creating access points to information. In latter, social paths serve as channels for rumours as well as expertise. Key problems include identification and disambiguation of individuals followed by difficulties of tracking the social connections. With semantic web, these aspects can be approached simultaneously. In this paper, we define a simple ontology for describing people and organizations. The model is based on FOAF and other existing vocabularies. We also demonstrate search and visualization tools for finding people.

1 Introduction

Social connections have been show to play an important role in getting the needed information. Granovetter (1973) argued that "weak" ties are most important in spreading information. (By a weak tie Granovetter means acquaintance like an old friend form school or work etc.) For example, most of "blue collar" jobs are shown to be passed through weak ties.

The web offers powerful tools for utilizing social connections (e.g. social networking sites like Facebook[1], Orkut[2] or Linked[3]). Machine driven mining is also been researched. Mika (2005); Aleman-Meza et al. (2007) have tried to build a kind of "who-is-who" index by crawling web pages, publications, emails etc.

Cross referencing and disambiguation has been long studied in library environment, where authors of similar name and documents with identical title are common. Authority control is a term that is used by library and information scientists to describe the methods for handling these problems.

Typical solution is to build an "authorized record" for each document and actor (person, group or organization). The record contains titles (and possibly their sources) and cross references. The following example is from a requirements document written by Functional Requirements and Numbering of Authority Records (FRANAR)[4] working group.

```
Authorized heading:
        Conti, House of
See also references:
        >> Bourbon, House of
        >> Cond, House of
See also reference tracings:
        << Bourbon, House of
        << Cond, House of
Cataloguers note:
        The House of Conti is a
        junior branch of the
        House of Bourbon-Cond.  Grand
        encyclopdie (Conti (maison de)).
```

Automatic tools for authority control include clustering French et al. (2000) and other name matching algorithms such as Galvez and Moya-Anegon (2007); Borgman and Siegfriend (1992).

Although authority control does not directly relate to social networking, one could use the rigorous methods for modelling entities and their connections. Name recognition and matching algorithms could also be useful e.g. in web crawler mining social networks. One example of a good social site with poor authority control is Last.fm[5] (problems date back to ambigous ID3 tags used in mp3s). In Figure 1 artists with same name are mixed. Also transliterations and other variations on names are not taken into account.

2 Actor Ontology

Our system includes extensive information about artists based on the Union List of Artist Names (ULAN)[6] vocabulary. ULAN consists of over

[1] http://www.facebook.com/
[2] http://www.orkut.com/
[3] http://www.linkedin.com/
[4] http://www.ifla.org/VII/d4/wg-franar.htm
[5] http://last.fm/
[6] http://www.getty.edu/research/conducting_research/vocabularies/ulan/

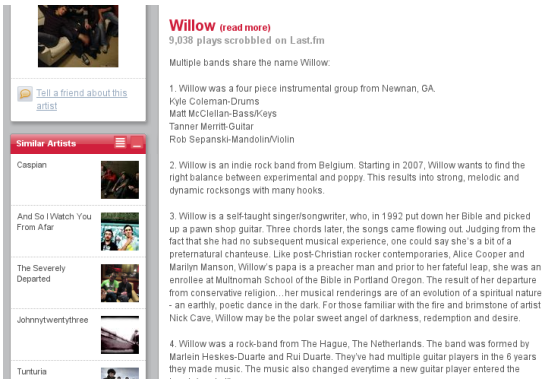

Figure 1: In Last.fm, using the name as an unique ID is causing problems. It is impossible to know, to which one of the four Willows the "Similar Artists" -recommendations are directed. Probably the recommendations are built to match the composition of these bands, and as such they might not match any of the Willows individually.

120,000 individuals and corporate bodies of art historical insignificance. In addition, data set includes comprehensive information about relationships between actors. As a strong authority record, ULAN contains over 300,000 names (Figure 2 shows an example of ULAN record). ULAN data was converted to ontological format using XSL-transformations.

Figure 2: Different names of Finnish artist Gallen-Kallela displayed on ULAN web site.

The model for our actor ontology is based on FOAF[7], Relationship[8] and BIO[9] vocabularies. Additional properties were added for roles and nationalities described in ULAN. In following example, a (non-ULAN) person is presented in RDF[10] with FOAF and other vocabularies.

```
<foaf:Person rdf:about=
  "http://www.yso.fi/onto/toimo/p12">
  <foaf:name>Jussi Kurki</foaf:name>
  <foaf:mbox>jussi.kurki@tkk.fi</foaf:mbox>
  <foaf:homepage rdf:resource=
    "http://www.seco.hut.fi/u/jhkurki"/>

  <bio:olb>
   Finnish student and research assistant
  </bio:olb>
  <bio:keywords>
    semantic web, computer science
  </bio:keywords>
  <bio:event>
    <bio:Birth>
      <bio:date>1982</bio:date>
      <bio:place>Helsinki</bio:place>
    </bio:Birth>
  </bio:event>

  <rel:worksWith rdf:resource=
   "http://www.yso.fi/onto/toimo/p23"/>
  <rel:worksWith rdf:resource=
   "http://www.yso.fi/onto/toimo/p61"/>
</foaf:Person>
```

In FOAF, the idea is to avoid global IDs e.g. URIs. Instead, person or group is identified by a set of unique properties like email or address. The process of merging data from different sources is called Smushing[11].

In actor ontology, we are indeed using URIs. To help resolving URIs, we have built a service called ONKI People which carries a similar idea that of ONKI Komulainen et al. (2005). ONKI People is a centralized repository of persons and organizations. It offers services for searching as well as disambiguating people.

3 ONKI People

Key features of ONKI People are multifaceted search component (Figure 3) and graph visualizer component (Figure 4). Search starts when user types one or more keywords to the search box and hits enter.

If user clicks an actor from the results list, the social circle of that actor is displayed. From the graph, user can further click any neighbours to see their social graphs. Graphs are rendered as SVG[12] images.

[7] http://xmlns.com/foaf/spec/
[8] http://vocab.org/relationship/
[9] http://vocab.org/bio/0.1/
[10] http://www.w3.org/RDF/
[11] http://wiki.foaf-project.org/Smushing
[12] http://www.w3.org/Graphics/SVG/About

Nodes are positioned by a simple algorithm which places direct contacts around the actor, friends of friends to the second level and so on.

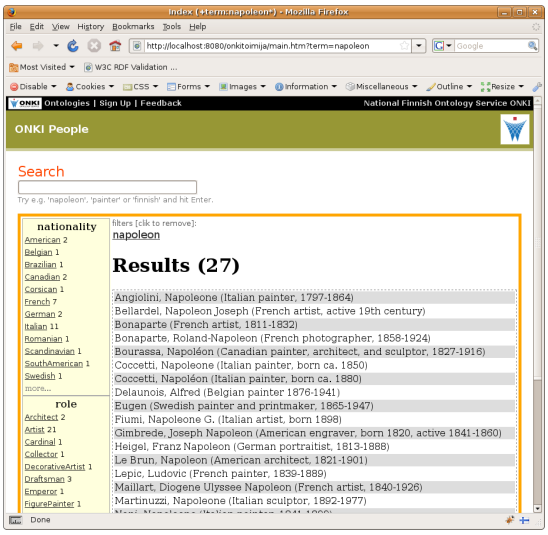

Figure 3: ONKI People showing the search results for keyword ”napoleon”.

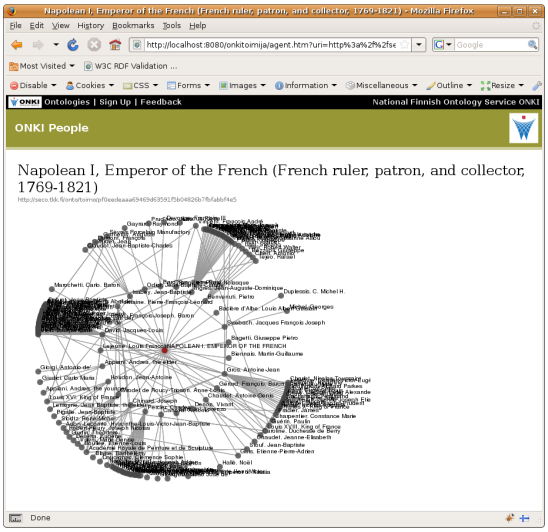

Figure 4: Displaying the social circle of Napoleon I in ONKI People.

ONKI People conforms also to the generic ONKI interface Viljanen et al. (2008) and can be published as a mash-up component using DWR[13]. Other machine interfaces, such as web services, could be easily added.

[13]http://directwebremoting.org/

ONKI People was implemented in Java on top of Spring framework[14]. Application follows Model-View-Controller (MVC) pattern where display logic is separated from the data model. As a view layer, JSP[15] and XSLT[16] were used. The search is backed by Lucene[17] index. In visualizer component, SVG graphs are rendered directly to HTTP-response to avoid the need of caching and disk operations. Other optimizations include compression of HTTP packets for faster page load times.

4 Relational Search

Semantic association identification has been studied in national security applications Sheth et al. (2005). We have built a system for searching semantic relations between persons. We have applied this notion to be called *relational semantic search* Kurki and Hyvnen (2007). (Similar work has been done in MultimediaN[18] portal.)

The idea is to make it possible for the end-user to formulate queries such as ”How is X related to Y” by selecting the end-point resources. The result is a set of semantic connection paths between X and Y.

For example, in Figure 5 the user has specified two historical persons, the Finnish artist Akseli Gallen-Kallela (1865–1931) and the French emperor Napoleon I (1769–1821) in a prototype of the portal Culturesampo Hyvönen et al. (2006). The system has discovered an association between the persons based on a chain of eight patronWas, teacherOf, and studentOf relations.

Relational search is done breath-first and even the longest paths (about 12 steps) can be found in less than half a second. This is explained partly by the structure of ULAN data. The graph has a strongly connected component of about 12000 actors containing central artists, such as Picasso and Donatello. At the same time, thousands of others, especially contemporary artists, don’t have any contacts in the underlying RDF graph.

The implementation was done in Java. A memory-based graph was built from the data and the graph was stored as adjacency list. To minimize memory consumption, graph node has only minimal set of fields: an id and a list of children. At this point, all relationships are basically reduced to ”knows” and all data is

[14]http://springframework.org/
[15]http://java.sun.com/products/jsp/
[16]www.w3.org/TR/xslt
[17]http://lucene.apache.org/
[18]http://e-culture.multimedian.nl/demo/search

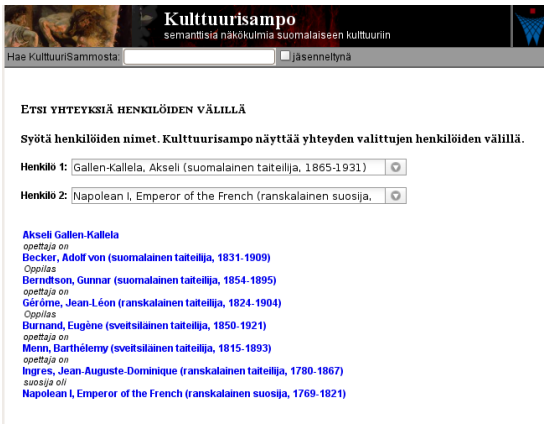

Figure 5: Relational search in Culturesampo using the ULAN vocabulary.

reduced to URI. Serialized to disk, the whole graph takes about 10MB of memory.

Though breadth-first search expands exponentially, it visits each node once at maximum. Search is obviously bounded by the size of the network and is thus $0(n)$.

5 Conclusions and future work

Social sites are gaining popularity as a way to find and access information. To fully enable social networking (and other linkage), identification and disambiguating should be handled better. Currently, it is difficult to combine knowledge from different sources. Even if the service providers agreed to it, different systems are using different formats for profiles. In addition, many sites use own local IDs for users (though recently an unified ID is been developed[19]).

A global search and ID repository could be handled with a help of service such as ONKI People, presented in this paper. To fully test this kind of functionality, user should be able to add and edit his or her own information.

Other possibility is to forget global IDs and centralized services – like FOAF is doing. Person writes and hosts his or her own profile. Social connections and other information identifies the person. One problem is that this requires some knowledge and effort from the user. Search is also difficult if there is no global index or structure on profiles.

To data annotators, such as librarians describing books or bloggers referring to people, ONKI People might be useful. Wikipedia, for example, already builds a record of people, and bloggers use wikipedia links to annotate people.

As shown, unified identifiers enable interesting services, such as relational search. As a part of semantic web, actors also link to other resources such as documents and pieces of art. This is been tested in Culturesampo Hyvönen et al. (2008). In future, we are planning on implementing a general relational search where the user can search connections between arbitrary resources.

Acknowledgements

This research was part of the National Finnish Ontology Project (FinnONTO) 2003-2007[20], funded mainly by The National Technology Agency (Tekes) and a consortium of 36 companies and public organisations. The work continues in FinnONTO 2.0 (2008-2009) project.

References

B. Aleman-Meza, U. Bojars, H. Boley, J. Breslin, M. Mochol, L. Nixon, A. Polleres, and A. Zhdanova. Combining rdf vocabularies for expert finding. In Enrico Franconi, Michael Kifer, and Wolfgang May, editors, *ESWC*, volume 4519 of *Lecture Notes in Computer Science*, pages 235–250. Springer, 2007.

C. Borgman and S. Siegfriend. Getty's synoname and its cousins: A survey of applications of personal name-matching algorithms. *Journal of the American Society for Information Science and Technology*, 43(7):459–476, 1992.

J. French, A. Powell, and E. Schulman. Using clustering strategies for creating authority files. *Journal of the American Society for Information Science*, 51(8):774–786, jun 2000.

C. Galvez and F Moya-Anegon. Approximate personal name-matching through finite-state graphs. *Journal of the American Society for Information Science and Technology*, 58(13):1960–1976, 2007.

M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–80, 1973.

Eero Hyvönen, Tuukka Ruotsalo, Thomas Häggström, Mirva Salminen, Miikka Junnila, Mikko Virkkilä, Mikko Haaramo, Eetu

[19] http://openid.net/

[20] http://www.seco.tkk.fi/projects/finnonto/

Mäkelä, Tomi Kauppinen, and Kim Viljanen. Culturesampo–finnish culture on the semantic web: The vision and first results. In *Developments in Artificial Intelligence and the Semantic Web - Proceedings of the 12th Finnish AI Conference STeP 2006*, October 26-27 2006.

Eero Hyvönen, Eetu Mäkelä, Tuukka Ruotsalo, Tomi Kauppinen, Olli Alm, Jussi Kurki, Joeli Takala, Kimmo Puputti, and Heini Kuittinen. Culturesampo–finnish culture on the semantic web. In *Posters of the 5th European Semantic Web Conference 2008 (ESWC 2008), Tenerife, Spain*, June 1-5 2008.

Ville Komulainen, Arttu Valo, and Eero Hyvönen. A collaborative ontology development and service framework ONKI. In *Proceeding of ESWC 2005, poster papers*. Springer, 2005.

Jussi Kurki and Eero Hyvnen. Relational semantic search: Searching social paths on the semantic web. In *Poster Proceedings of the International Semantic Web Conference (ISWC 2007), Busan, Korea*, Nov 2007.

Peter Mika. Flink: Semantic web technology for the extraction and analysis of social networks. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):211–223, October 2005.

Amit Sheth, Boanerges Aleman-Meza, I. Budak Arpinar, Clemens Bertram, Yashodhan Warke, Cartic Ramakrishnan, Chris Halaschek, Kemafor Anyanwu, David Avant, F. Sena Arpinar, and Krys Kochut. Semantic association identification and knowledge discovery for national security applications. *Journal of Database Management on Database Technology*, 16(1):33–53, Jan–March 2005.

Kim Viljanen, Jouni Tuominen, and Eero Hyvönen. Publishing and using ontologies as mash-up services. In *Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW2008), 5th European Semantic Web Conference 2008 (ESWC 2008)*, June 1-5 2008.