A Distributed Entity Directory*

Fausto Giunchiglia and Alethia Hume

Department of Information Engineering and Computer Science
University of Trento, Italy
{fausto,hume}@disi.unitn.it
http://www.disi.unitn.it

Abstract. We see the local content from peers organized in directories (i.e., on local ordered lists) containing local representations of entities from the real world (e.g., persons, locations, events). Different local representations can give different "versions" of the same real world entity and use different names to refer to it (e.g., Fausto Giunchiglia, Giunchiglia F., Prof. Giunchiglia). Although the names used in these directories connect data that could complement each other, there are no links that allow peers to share and search across them. We propose a Distributed Directory of Entities that makes explicit these connecting links and allows peers to: (i) maintain their data locally and (ii) find the different versions of a real world entity based on any name used in the network. The model we present exploits the name as the central (multi-value) attribute of entities and aims to convince readers of the importance of such names in a peer-to-peer scenario.

1 A Distributed Entity Directory

We see the internet as a network of peers maintaining local directories of entities that are interconnected. An example of this is shown in the first part of Figure 1, where different local representations are seen as pieces of information about a particular entity that are stored in a distributed manner in the network. Entities can be of different types, they have a name, and are described by attributes (e.g., latitude-longitude, size, birth date). However, the names of entities play a key role as identifiers, which are used to distinguish them from others (e.g., F. Giunchiglia, Italy, University of Trento) and behave similarly to keywords. Moreover, real world entities can be called by multiple names as a consequence of different types of variations and errors (e.g., format variations, partial and full translations, misspellings and pseudonyms).

We formalize these characteristics in a Distributed Directory[1] that distinguishes between a Digital Entity (DE) and a Real World Entity (WE):

$$DE = \langle URL, \{N\} \rangle \qquad (1) \qquad\qquad WE = \langle URI, \{URL\} \rangle \qquad (2)$$

A DE is a local representation of an entity that exists in the real world, while

* This work has been partially supported by the EU-FET grant SmartSociety 600854.

[1] The interested reader can find an extended version of the approach in [1].

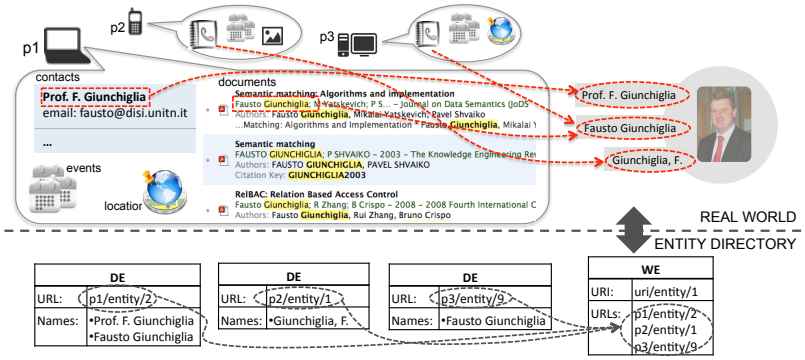

Fig. 1. Example of a network of interconnected directories.

a WE represents the real world entity (modeled as a class of DEs). A URL is used to uniquely identify a DE and it can be dereferenced to obtain the local description based on attributes; $\{N\}$ is the set of names used in DEs to refer to a WE. A URI is used to uniquely identify a WE and a non-empty set $\{URL\}$ contains the identifiers of different DEs that describe WE (see Figure 1).

The notions introduced in (1) and (2) define the *many-to-many* relation between $Names$ and WEs. Different names can be used in DEs to identify the same WE. At the same time, the names used in DEs that refer to different WEs can overlap. We exploit these notions by building two indexes:

1. A *DEindex* maps WEs (i.e., $URIs$) to DEs (i.e., $URLs$) containing the different versions that locally represent them.
2. A *WEindex* maps the *names* (given in DEs) to WEs (i.e., $URIs$) that are candidates to be called by such names.

The two indexes allow finding all the different versions of an entity based on any name used in the network to refer to it. The fact that the indexes contain only identifiers (i.e., names, $URIs$ and $URLs$) allows peers to always maintain their attribute-based descriptions locally. Furthermore, in our approach the indexes are stored using a Distributed Hash Table (DHT) built on top of the same network of peers that share and consume the data, which gives us scalability.

We evaluated the approach on networks of 50, 100, and 150 peers running on PlanetLab. The average query time (as a measure of the performance) is 2.7 seconds for the different network sizes. We also noticed that more than 55% of the queries are answered in less than a second, almost 70% in less than 2 seconds and only 9% of queries takes more than 5 seconds to be answered. We believe these results are promising in terms of scalability because they show that performance can be stable with the network growth.

References

1. Giunchiglia, F., Hume, A.: A distributed directory system. Technical Report DISI-13-005, University of Trento, Italy (2013)