Chapter 8
Data Linking for the Semantic Web

Alfio Ferraram
Università degli Studi di Milano, Italy

Andriy Nikolov
The Open University, UK

François Scharffe
University of Montpellier, France

ABSTRACT
By specifying that published datasets must link to other existing datasets, the 4th linked data principle ensures a Web of data and not just a set of unconnected data islands. The authors propose in this paper the term data linking to name the problem of finding equivalent resources on the Web of linked data. In order to perform data linking, many techniques were developed, finding their roots in statistics, database, natural language processing and graph theory. The authors begin this paper by providing background information and terminological clarifications related to data linking. Then a comprehensive survey over the various techniques available for data linking is provided. These techniques are classified along the three criteria of granularity, type of evidence, and source of the evidence. Finally, the authors survey eleven recent tools performing data linking and we classify them according to the surveyed techniques.

INTRODUCTION
In the Semantic Web and in the Web in general, a fundamental problem is the comparison and matching of data and the capability of resolving the multiplicity of data references to the same real-world objects, by defining correspondences among data in form of data links. The data linking task is becoming more and more important as the number of structured and semistructured data available on the Web is growing. The transformation of the Web from a “Web of documents” into a “Web of data”, as well as the availability of large collections of sensor generated data (“internet of
things”), is leading to a new generation of Web applications based on the integration of both data and services. At the same time, new data are published every day out of user generated contents and public Web sites.

In general terms, data linking is the task of determining whether two object descriptions can be linked one to the other to represent the fact that they refer to the same real-world object in a given domain or the fact that some kind of relation holds between them. Quite often, this task is performed on the basis of the evaluation of the degree of similarity among different data instances describing real-world objects across heterogeneous data sources, under the assumption that the higher is the similarity between two data descriptions, the higher is the probability that the two descriptions actually refer to the same object. From an operative point of view, data linking includes also the task of defining methods, techniques and (semi-) automated tools for performing the similarity evaluation task. We call this specific subtask instance matching.

In this context, one of the most important initiatives in the Semantic Web field is the Linked Open Data project, which promotes the idea of improving interoperability and aggregation among large data collections already available on the Web (Berners-Lee et al., 2008). The example of Linked Data shows how the data linking task is crucial on the Web nowadays. Fortunately, there is a lot of work done in other fields that can be used to provide solutions, methods, and techniques to address data linking on the Semantic Web. Moreover, there are important works describing research fields that are very close to data linking (Euzenat & Shvaiko, 2007; Koudas et al., 2006; Bleiholder & Naumann, 2009). However, there are specific features of semantic data which require specific solutions both in terms of new techniques and in terms of original combinations of techniques that have been originally proposed in different fields. For example, on one side, data linking requires dealing with the semantic complexity which is typical of ontology matching, but, on the other side, the large amount of data available on the “Web of data” requires to deal with scalability issues that are typical of record linkage. This situation leads to the development of new approaches, addressing problems that are typical of the data linking field on the Semantic Web. In this paper, we provide a general definition of this field, in order to underline problems and to describe solutions. In particular, in the next section, we will better define the data linking problem, by discussing also the terminology used in the field. After that, we present the main families of techniques proposed for the most relevant subtask of data linking, that is instance matching, then we survey the most relevant tools in the field by comparing them. Finally, we discuss the main open problems and directions of work in the field.

**PROBLEM FORMULATION**

Data linking can be formalized as an operation which takes two collections of data as input and produces a collection of mappings between entities of the two collections as output. Mappings denote binary relations between entities corresponding semantically one to another. The data linking task is articulated in steps as shown in Figure 1.

The input of the process is given by one or more datasets. Each dataset is a collection of data representing object descriptions to be linked. The output of the process is a mapping set that is a collection of binary relations (usually referred as mappings or links) between the object descriptions in the input dataset(s). The data linking task also involves a user, who has the responsibility of configuring the process and, optionally, interacting with the predicate selection, pre-processing, matching and post-processing steps in case of a semi-automatic procedure. Another optional component is provided by external resources such as lexical databases and thesauri, reasoning systems or pre-defined mappings that can be used both in the matching and in the post-processing steps as a support for the main process. A typical