

# Entity Resolution in the Web of Data

# Synthesis Lectures on the Semantic Web

## Editor

**Ying Ding, *Indiana University***

**Paul Groth, *Elsevier Labs***

*Synthesis Lectures on the Semantic Web: Theory and Application* is edited by Ying Ding of Indiana University and Paul Groth of Elsevier Labs. Whether you call it the Semantic Web, Linked Data, or Web 3.0, a new generation of Web technologies is offering major advances in the evolution of the World Wide Web. As the first generation of this technology transitions out of the laboratory, new research is exploring how the growing Web of Data will change our world. While topics such as ontology-building and logics remain vital, new areas such as the use of semantics in Web search, the linking and use of open data on the Web, and future applications that will be supported by these technologies are becoming important research areas in their own right. Whether they be scientists, engineers or practitioners, Web users increasingly need to understand not just the new technologies of the Semantic Web, but to understand the principles by which those technologies work, and the best practices for assembling systems that integrate the different languages, resources, and functionalities that will be important in keeping the Web the rapidly expanding, and constantly changing, information space that has changed our lives.

Topics to be included:

- Semantic Web Principles from linked-data to ontology design
- Key Semantic Web technologies and algorithms
- Semantic Search and language technologies
- The Emerging "Web of Data" and its use in industry, government and university applications
- Trust, Social networking and collaboration technologies for the Semantic Web
- The economics of Semantic Web application adoption and use
- Publishing and Science on the Semantic Web
- Semantic Web in health care and life sciences

## Entity Resolution in the Web of Data

Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis

2015

## Library Linked Data in the Cloud: OCLC's Experiments with New Models of Resource Description

Carol Jean Godby, Shenghui Wang, and Jeffrey K. Mixer  
2015

## Semantic Mining of Social Networks

Jie Tang and Juanzi Li  
2015

## Social Semantic Web Mining

Tope Omitola, Sebastián A. Ríos, and John G. Breslin  
2015

## Semantic Breakthrough in Drug Discovery

Bin Chen, Huijun Wang, Ying Ding, and David Wild  
2014

## Semantics in Mobile Sensing

Zhixian Yan and Dipanjan Chakraborty  
2014

## Provenance: An Introduction to PROV

Luc Moreau and Paul Groth  
2013

## Resource-Oriented Architecture Patterns for Webs of Data

Brian Sletten  
2013

## Aaron Swartz's A Programmable Web: An Unfinished Work

Aaron Swartz  
2013

## Incentive-Centric Semantic Web Application Engineering

Elena Simperl, Roberta Cuel, and Martin Stein  
2013

## Publishing and Using Cultural Heritage Linked Data on the Semantic Web

Eero Hyvönen  
2012

## VIVO: A Semantic Approach to Scholarly Networking and Discovery

Katy Börner, Michael Conlon, Jon Corson-Rikert, and Ying Ding  
2012

## Linked Data: Evolving the Web into a Global Data Space

Tom Heath and Christian Bizer  
2011

© Springer Nature Switzerland AG 2022  
Reprint of original edition © Morgan & Claypool 2015

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Entity Resolution in the Web of Data

Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis

ISBN: 978-3-031-79467-4      paperback

ISBN: 978-3-031-79468-1      ebook

DOI 10.1007/978-3-031-79468-1

A Publication in the Springer series  
*SYNTHESIS LECTURES ON THE SEMANTIC WEB*

Lecture #13

Series Editors: Ying Ding, *Indiana University*

Paul Groth, *Elsevier Labs*

Founding Editor Emeritus: James Hendler, *Rensselaer Polytechnic Institute*

Series ISSN

Print 2160-4711    Electronic 2160-472X

# Entity Resolution in the Web of Data

Vassilis Christophides

University of Crete, Greece  
INRIA, France

Vasilis Efthymiou

University of Crete, Greece  
ICS-FORTH, Greece

Kostas Stefanidis

ICS-FORTH, Greece

*SYNTHESIS LECTURES ON THE SEMANTIC WEB #13*

## ABSTRACT

In recent years, several knowledge bases have been built to enable large-scale knowledge sharing, but also an entity-centric Web search, mixing both structured data and text querying. These knowledge bases offer machine-readable descriptions of real-world entities, e.g., persons, places, published on the Web as Linked Data. However, due to the different information extraction tools and curation policies employed by knowledge bases, multiple, *complementary* and sometimes *conflicting* descriptions of the same real-world entities may be provided. Entity resolution aims to identify different descriptions that refer to the same entity appearing either within or across knowledge bases.

The objective of this book is to present the *new entity resolution challenges* stemming from the *openness* of the Web of data in describing entities by an unbounded number of knowledge bases, the *semantic and structural diversity* of the descriptions provided across domains even for the same real-world entities, as well as the *autonomy* of knowledge bases in terms of adopted processes for creating and curating entity descriptions. The *scale*, *diversity*, and *graph structuring* of entity descriptions in the Web of data essentially challenge how two descriptions can be effectively compared for similarity, but also how resolution algorithms can efficiently avoid examining pairwise all descriptions.

The book covers a wide spectrum of entity resolution issues at the Web scale, including basic concepts and data structures, main resolution tasks and workflows, as well as state-of-the-art algorithmic techniques and experimental trade-offs.

## KEYWORDS

entity resolution, Web of data

# Contents

	<b>List of Figures</b> .....	<b>ix</b>
	<b>List of Tables</b> .....	<b>xi</b>
	<b>Preface</b> .....	<b>xiii</b>
	<b>Acknowledgments</b> .....	<b>xv</b>
<b>1</b>	<b>Web of Data: Describing and Linking Entities</b> .....	<b>1</b>
<b>2</b>	<b>Matching and Resolving Entities</b> .....	<b>17</b>
2.1	The Problem of Entity Resolution .....	17
2.2	Similarity Functions .....	22
2.2.1	Content-based Similarity Functions .....	25
2.2.2	Relational Similarity Functions .....	28
2.2.3	Approximations of Similarity Functions .....	33
2.3	Discussion .....	37
<b>3</b>	<b>Blocking</b> .....	<b>39</b>
3.1	The Problem of Entity Blocking .....	39
3.2	Blocking in Traditional Data Warehouses .....	40
3.3	Blocking in the Web of Data .....	42
3.4	Block Post-processing Methods .....	49
3.5	Discussion .....	52
<b>4</b>	<b>Iterative Entity Resolution</b> .....	<b>55</b>
4.1	The Problem of Iterative Entity Resolution .....	55
4.2	Merging-based Iterative Entity Resolution .....	57
4.3	Relationship-based Iterative Entity Resolution .....	60
4.4	Iterative Blocking .....	65
4.5	Incremental Entity Resolution .....	68
4.6	Progressive Entity Resolution .....	68
4.7	Discussion .....	71

<b>5</b>	<b>Experimental Evaluation of Blocking Algorithms</b>	<b>73</b>
5.1	Datasets	73
5.2	Measures	77
5.3	Quality Results	79
5.3.1	Identified Matches (TPs)	79
5.3.2	Missed Matches (FNs)	82
5.3.3	Non-matches (FPs and TNs)	83
5.4	Performance Results	84
5.5	Different Types of Links	85
5.6	Lessons Learned	85
<b>6</b>	<b>Conclusions</b>	<b>87</b>
	<b>Bibliography</b>	<b>91</b>
	<b>Authors' Biographies</b>	<b>105</b>



# List of Figures

1.1	A part of the Web of data, extracted from three knowledge bases. . . . .	2
1.2	The Linked Open Data cloud of Web KBs. . . . .	9
1.3	In- (top) and out-degree (bottom) distributions of different categories of datasets. . . . .	10
1.4	Searching and recommending entities related to “Stanley Kubrick.” . . . .	15
2.1	Multiple entity descriptions. . . . .	18
2.2	Entity resolution process. . . . .	21
2.3	Ideal and realistic similarity functions in entity resolution. . . . .	23
2.4	An example of a relational star schema. . . . .	28
2.5	An example RDF graph, used to evaluate $LINDA_{sim}$ . . . . .	31
2.6	An example RDF graph, used to evaluate $sim_{ov}$ . . . . .	32
2.7	An example characteristic matrix (a), three random permutations $h1, h2, h3$ of the rows of this matrix (b), and the resulting minhash signature matrix (c). . . . .	34
2.8	The S-curve. . . . .	36
3.1	A set of entity descriptions. . . . .	42
3.2	Token blocking example. . . . .	43
3.3	Attribute clustering blocking example: (a) most similar attribute-pairs, (b) attribute clusters, (c) generated blocks. . . . .	45
3.4	Prefix-infix(-suffix) blocking example: (a) the input entity collection, (b) URI identifiers of the descriptions, (c) generated blocks. . . . .	46
3.5	The blocks generated by a set-similarity join method for the descriptions of Figure 3.1. . . . .	48
3.6	The blocks of Figure 3.2 in ascending order of size (top) and the corresponding entity index (bottom). . . . .	51
3.7	Meta-blocking example: (a) depicts a blocking graph, which is pruned (b), to discard unnecessary comparisons. . . . .	52

4.1	A merging-based iterative ER example (a) and a relationship-based iterative ER example (b). . . . .	56
4.2	The execution of R-Swoosh for Example 4.3. . . . .	59
4.3	Two different descriptions of the movie <i>A Clockwork Orange</i> and its cast in XML. . . . .	61
4.4	An entity graph used by collective entity resolution. . . . .	62
4.5	An execution example of LINDA. (a) PQ initialization, (b) PQ update, (c) new matches are found, (d) distributed version. . . . .	64
4.6	An example showing the process of iterative blocking. . . . .	66
4.7	The general structure of HARRA. . . . .	67
4.8	Progressive ER process. . . . .	69
4.9	A progressive ER algorithm <i>A</i> , compared to a typical ER algorithm <i>B</i> . . . . .	71
5.1	Common tokens (top) and common tokens in common clusters (bottom) per entity description distributions for <i>D1–D6</i> . . . . .	82

# List of Tables

1.1	The RDF triples of the entities appearing on Figure 1.1 . . . . .	3
1.2	Comparison of knowledge bases . . . . .	5
1.3	Top-3 properties used by RDF links within each topical domain in the 2014 LOD cloud . . . . .	11
1.4	Top 10 KBs based on their number of incoming <i>owl:sameAs</i> links . . . . .	11
3.1	Criteria for placing descriptions in the same block . . . . .	53
3.2	Blocking approaches with respect to the redundancy attitude and algorithmic attitude . . . . .	53
5.1	Datasets characteristics . . . . .	75
5.2	Characteristics of datasets with different types of links to <i>BTC12DBpedia</i> . . . .	75
5.3	Entity collections characteristics . . . . .	76
5.4	Definitions for pairs of descriptions, based on whether they appear in a common block, or not . . . . .	77
5.5	Quality measures (the ideal value of each measure is in boldface) . . . . .	78
5.6	Statistics and evaluation of blocking methods . . . . .	81
5.7	Characteristics of the missed matches (false negatives) of token blocking . . . .	84
5.8	Recall of token blocking for the collections composed of datasets of Table 5.2 and <i>BTC12DBpedia</i> . . . . .	85

# Preface

Over the past decade, numerous knowledge bases (KBs) have been built to power a new generation of Web applications that provide *entity-centric search* and *recommendation services*. These KBs offer comprehensive, machine-readable descriptions of a large variety of real-world entities (e.g., persons, places, products, events) published on the Web as Linked Data (LD). Even when derived from the same data source (e.g., a Wikipedia entry), KBs such as DBpedia, YAGO2, or Freebase may provide multiple, non-identical descriptions for the same real-world entities. This is due to the different information extraction tools and curation policies employed by KBs, resulting in *complementary* and sometimes *conflicting* entity descriptions. Entity resolution (ER) aims to identify different descriptions that refer to the same real-world entity, and emerges as a central data-processing task for an *entity-centric organization* of Web data. ER is needed to enrich inter-linking of data elements describing entities, even by third parties, so that the Web of data can be accessed by machines as a *global data space* using standard languages, such as SPARQL. ER can also facilitate an automated KB construction by integrating entity descriptions from legacy KBs with Web content published as HTML documents.

ER has attracted significant attention from many researchers in information systems, database, and machine-learning communities. The objective of this lecture is to present the *new ER challenges* stemming from the Web *openness* in describing, by an unbounded number of KBs, a multitude of entity types across domains, as well as the *high heterogeneity* (semantic and structural) of descriptions, even for the same types of entities. The *scale*, *diversity*, and *graph structuring* of entity descriptions published according to the LD paradigm challenge the core ER tasks, namely, (i) how descriptions can be *effectively compared for similarity* and (ii) how resolution algorithms can *efficiently filter the candidate pairs* of descriptions that need to be compared.

In a multi-type and large-scale entity resolution, we need to examine whether two entity descriptions are *somehow* (or near) *similar* without resorting to domain-specific similarity functions and/or mapping rules. Furthermore, the resolution of some entity descriptions might influence the resolution of other neighborhood descriptions. This setting clearly goes beyond deduplication (or record linkage) of collections of descriptions, usually referring to a single entity type, that slightly differ only in their attribute values. It essentially requires leveraging similarity of descriptions both on their *content* and *structure*. It also forces us to revisit traditional ER workflows consisting of separate *indexing* (for pruning the number of candidate pairs) and *matching* (for resolving entity descriptions) phases.

This Synthesis lecture is intended to provide a starting point for researchers, students, and developers who are interested in a global view of the ER problem in the Web of data. Throughout the lecture, we present the basic concepts and resolution workflows, as well as state-of-the-art

indexing and matching techniques. We additionally survey new ER execution strategies (such as parallel/distributed and progressive strategies) to resolve, under specific efficiency or effectiveness constraints, very large collections of entity descriptions, eventually arriving in streams. We made an effort to define in a self-contained way the similarity measures and data structures involved in various algorithms along with representative examples. We finally provide an experimental evaluation of a large part of the presented techniques and explain the involved trade-offs for real KBs in the Web of data. Much of the material presented in this lecture has been used in graduate courses taught at the University of Crete, as well as in two recent tutorials at CIKM'13 and WWW'14.

Since ER is a specialized problem of Data Integration, our Synthesis lecture provides complementary material with other books in this research area. [Doan et al., 2012] focuses on models, languages, and architectures for Data Integration systems, as well as on techniques for rewriting and processing queries on top. It also covers machine learning techniques for inferring mappings/matchings between heterogeneous relational and Web data. [Dong and Srivastava, 2015] stresses the Data Integration challenges in the Big Data era. In particular, it details how well-known ER algorithms can benefit from parallel and distributed implementations, aiming to reduce the overall execution time of the entire ER process. The book also considers schema alignment, as well as techniques for linking text snippets with embedded attributes, to structured records. Record Linkage techniques and Deduplication techniques for traditional Data Warehouse settings have been the subject of numerous surveys and books, such as [Naumann and Herschel, 2010] and [Christen, 2012]. Finally readers are referred to [Abiteboul et al., 2011] for a comprehensive overview of languages and technologies involved in Web Data Management.

Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis  
June 2015

# Acknowledgments

Several people provided valuable support during the preparation of this book, without whose help the project could not have been satisfactorily and timely completed. We warmly thank Ying Ding and Paul Groth for inviting us to write this book and Michael Morgan for managing the entire publication process. Special thanks also go to Christian Bizer for constructive feedback during the review process of the book. We would also like to acknowledge our many collaborators who have influenced our thoughts and our understanding of this research area over the years, and the following projects for their support in our research efforts: EU FP7-ICT-2011-9 DIACHRON (Managing the Evolution and Preservation of the Data Web), EU FP7-PEOPLE-2013-IRSES SemData (Semantic Data Management), EU FP7-ICT-318552 IdeaGarden (An Interactive Learning Environment Fostering Creativity), and LoDGoV (Generate, Manage, Preserve, Share, and Protect Resources in the Web of Data) of the Research Programme ARISTEIA (EXCELLENCE), GSRT, Ministry of Education, Greece, and the European Regional Development Fund. Finally, we would like to thank the ~okeanos GRNET cloud service that is used in our experimental evaluation.

Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis  
June 2015