The wealth of freely available, structured information on the Web is constantly growing. Driving domains are public data from and about governments and administrations, scientific data, and data about media, such as articles, books and albums. In addition, general-purpose datasets, such as DBpedia and Freebase from the linked open data community, serve as a focal point for many data sets. Thus, it is possible to query or integrate data from multiple sources and create new, integrated data sets with added value.

Yet integration is far from simple: It happens at technical level by ingesting data in various formats, at structural level by providing a common ontology and mapping the data source structures to it, and at semantic level by linking multiple records about same real world entities and fusing these representations into a clean and consistent record. The talk highlights the extreme heterogeneity of web data and points to methods to overcome them including a multitude of tasks that must be completed: source selection to identify appropriate and high quality sources, data extraction to create structured data, scrubbing to standardize and clean data, entity matching to associate different occurrences of the same entity, and finally data transformation and data fusion to combine all data about an entity in a single, consistent representation.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous

General Terms
Algorithms, Management

Keywords
Linked Open Data, Data Cleansing, Data Integration

Copyright is held by the author/owner(s).

BEWEB '11, March 25, 2011, Uppsala, Sweden Copyright 2011 ACM 978-1-4503-0610-2/11/03 . . $10.00.