The Web has grown into a heterogeneous open data space of interlinked documents, tables, and databases. Analyzing these datasets in knowledge discovery processes requires careful data preparation, which usually takes 60% of the total time spent. Statistical data, which are data subject to analysis by statistical methods, show a number of problems making their preparation difficult: legacy and hardly accessible tabular formats, disturbing data errors, and non-reusable data curation procedures are some examples. In addition, historical statistical data are only partially preserved, and their harmonization is very poor. Today, these issues are only addressed with inefficient and non-reproducible data munging.

This thesis contributes solutions to these problems, taking advantage of Semantic Web technologies, and using the domain of Social and Economic History as a case study. Concretely, its contributions are:

- A thorough study of data preparation and integration requirements in Social Science and History
- A semi-automatic approach that combines human expertise with scalable automation to convert legacy statistical collections to 5-star Linked Data
- A novel metric to quantify the predictability of diachronic Web schemas
- A Web-friendly way of expressing statistical data validation procedures (or "edit rules") in SPARQL, the RDF query language
- An easy way of extending the functionality of SPARQL by leveraging federation

These contributions are a basis for bringing a more efficient data preparation and refinement to the Web in quantitative History, the Digital Humanities, and Science.

VU UNIVERSITY AMSTERDAM

eHumanities
Royal Netherlands Academy of Arts and Sciences

Data Archiving and Networked Services
DANS

REFINING STATISTICAL DATA ON THE WEB

ALBERT MEROÑO PEÑUELA

REFINING STATISTICAL DATA ON THE WEB

ALBERT MEROÑO PEÑUELA