E-Science-Initiative:
Virtuelle Forschungsumgebungen

Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

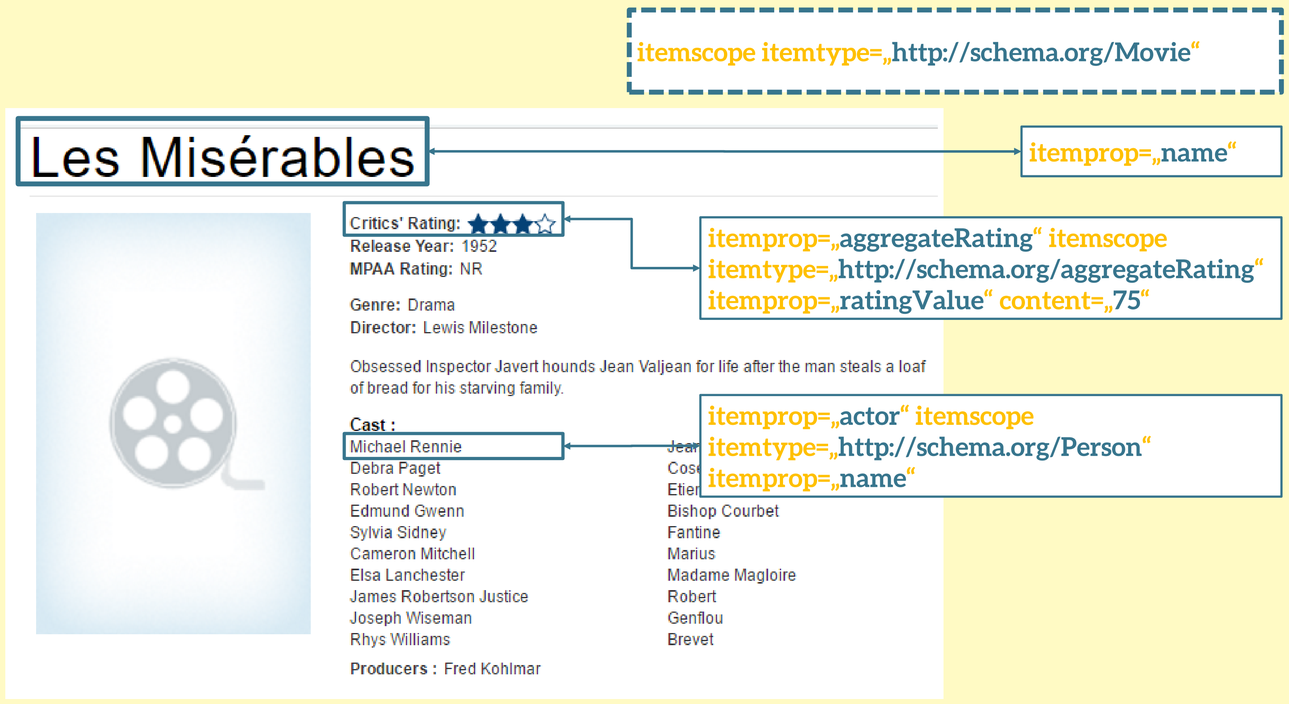
The Web Data Commons Structured Data Extraction

Anna Primpeli, Robert Meusel, Christian Bizer, Heiner Stuckenschmidt
Data and Web Science Group, University of Mannheim

Extracting structured data out of 3.1 billion web pages from the biggest public web corpus
- A use case of the ViCE project -

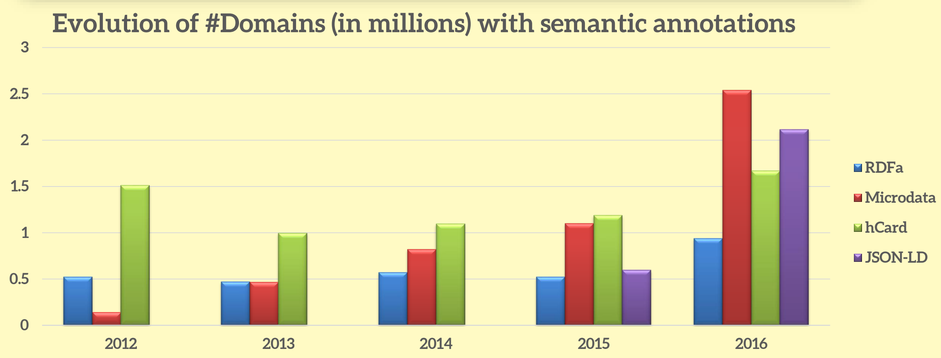
Structured Data on the Web

A Semantic Annotation Example



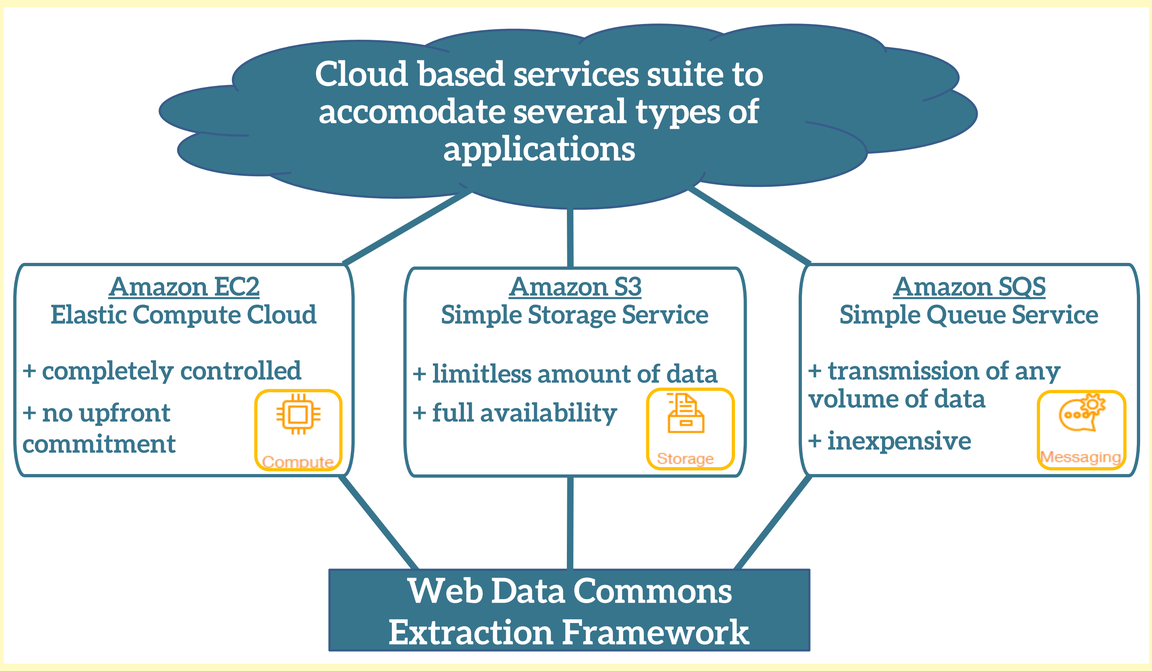
Common Topics

Evolution

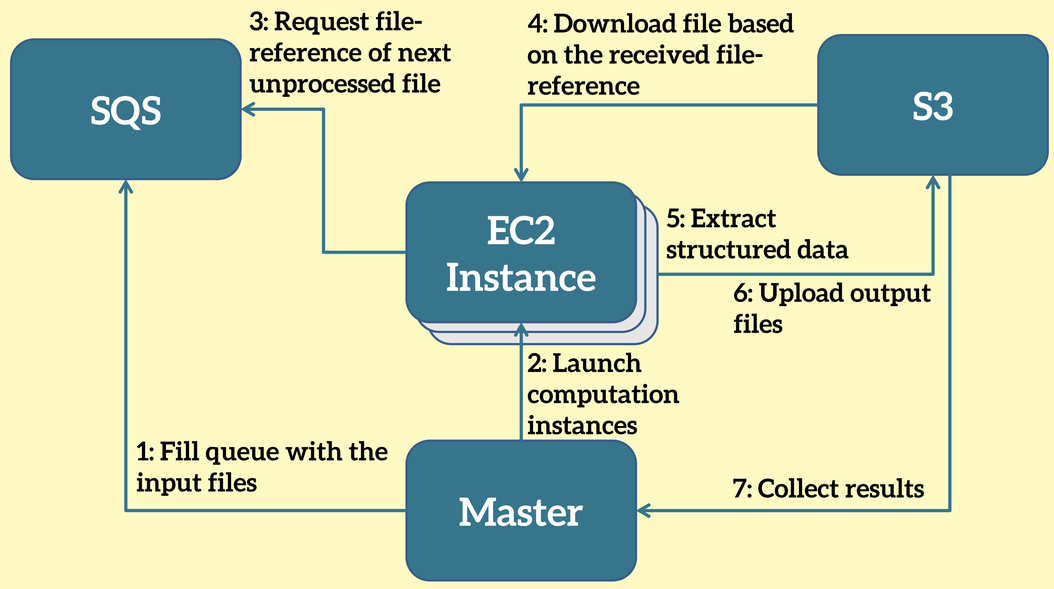


Cloud Based Extraction

Amazon Web Services



Extraction Framework Architecture



Web Data Commons, Requirements and Results

Web Data Commons

The Web Data Commons project extracts structured data from the Common Crawl, the largest web corpus available to the public, and provides the extracted data for public download. Different types of data have been generated since 2012 till today including the following:

RDFa, Microdata, and Microformat Data Sets
The extracted data as well as statistics about the deployment of different formats are published

Web Tables Corpus
Dataset containing 147 million relational web tables

Hyperlink Graph
Covers 3.5 billion web pages and 128 billion hyperlinks between these pages

WebIsA Database
Contains more than 400 million hypernymy relations

Product Data Corpus
Contains over 5.6 million product records retrieved from the most visited 32 shopping websites

Requirements

Processing
- 100 computation nodes
- 8 cores, 64-bit, 2.8 GHz
- 15 GB RAM
- 2x80 GB SSD

Data Storage
- 56 TB of input data from Common Crawl
- 1 TB of resulting data

Time & Cost
- 50 hours
- 650 $ total cost

Results

Input Data
56 TB
34 million Domains
3.1 billion URLs

Extracted Data
1 TB Extracted Data
5.6 million Domains
1.2 billion URLs

Extracted Data as NQuads
44 billion Triples
9.5 billion Entities

WDC Structured Data
http://webdatacommons.org/structureddata

UNIVERSITÄT MANNHEIM

ViCE
Virtual Open Science Collaboration Environment

Web Data Commons
http://webdatacommons.org

Das Projekt wird gefördert vom Ministerium für Wissenschaft, Forschung und Kunst, Baden Württemberg.