Roomba

Automatic Validation, Correction and Generation of Dataset Metadata Enhancing Dataset Search and Spam Detection

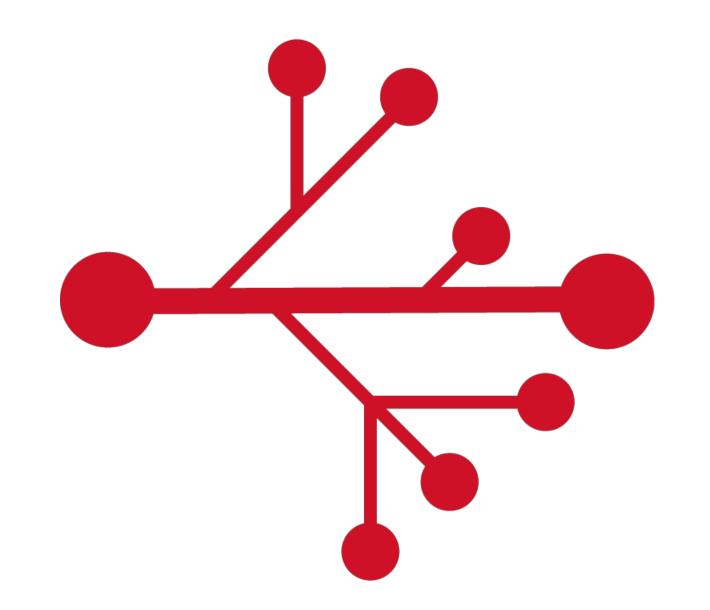




Ahmad Assaf+, Raphaël Troncy+ and Aline Senart‡

†EURECOM Sophia Antipolis, France firstName.lastName@eurecom.fr

‡ SAP Research, SAP Labs France SAS Mougins, France aline.senart@sap.com

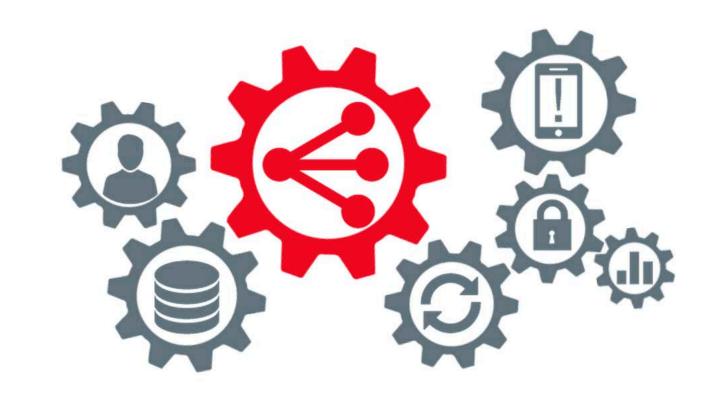


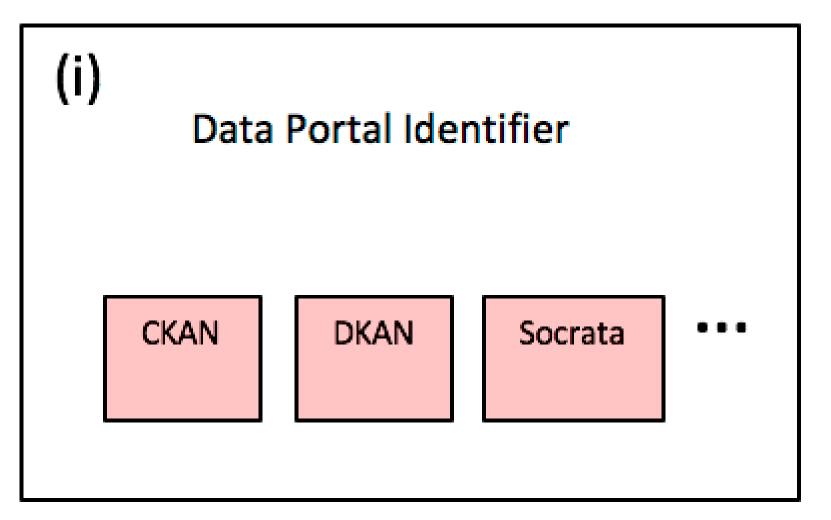
Data Portals

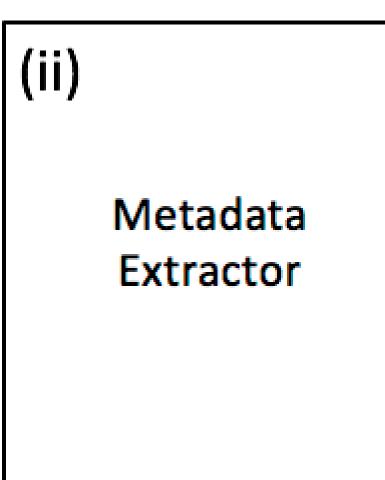
A curated collection of metadata about datasets providing discovery functionality to complement conventional browse-style catalogue interfaces

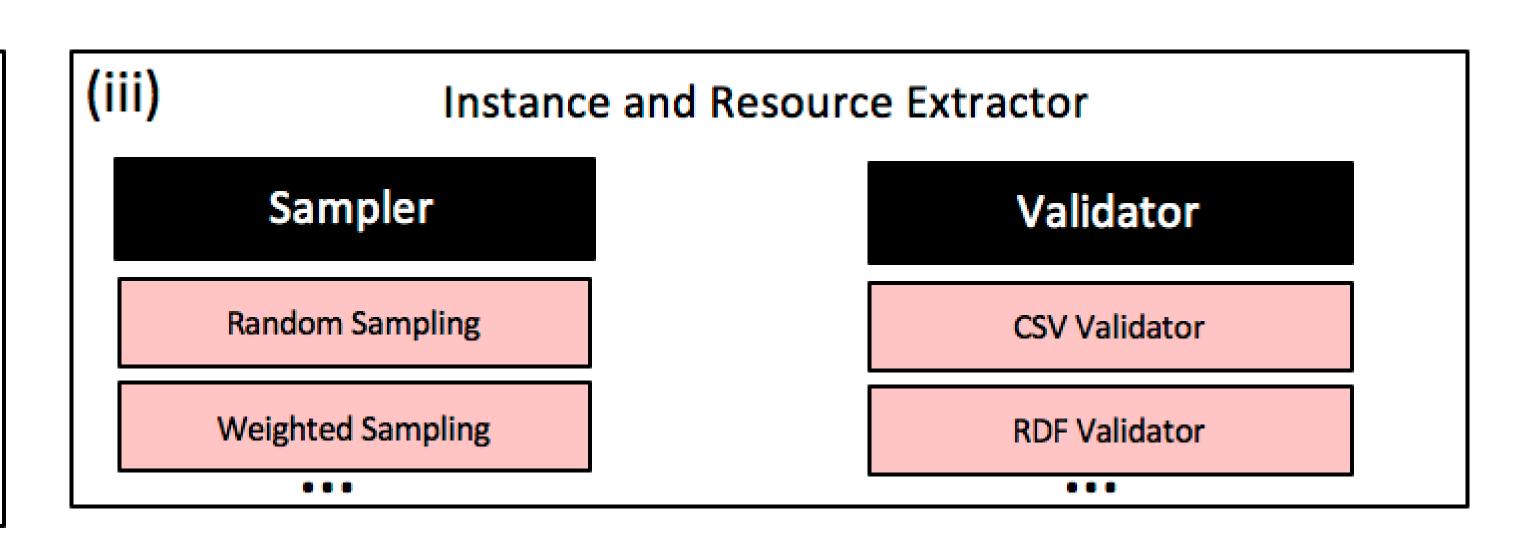


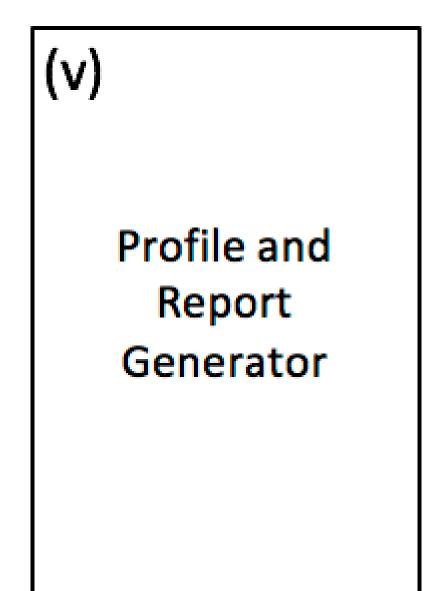


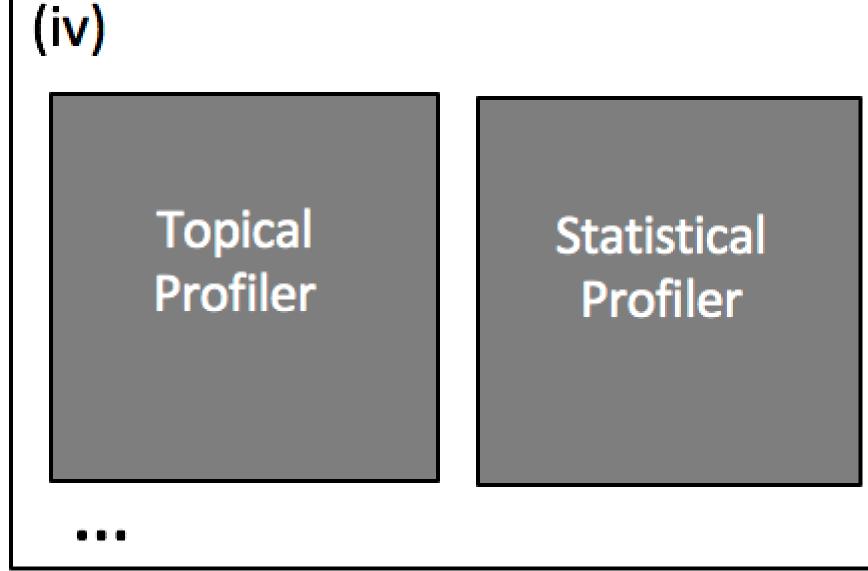


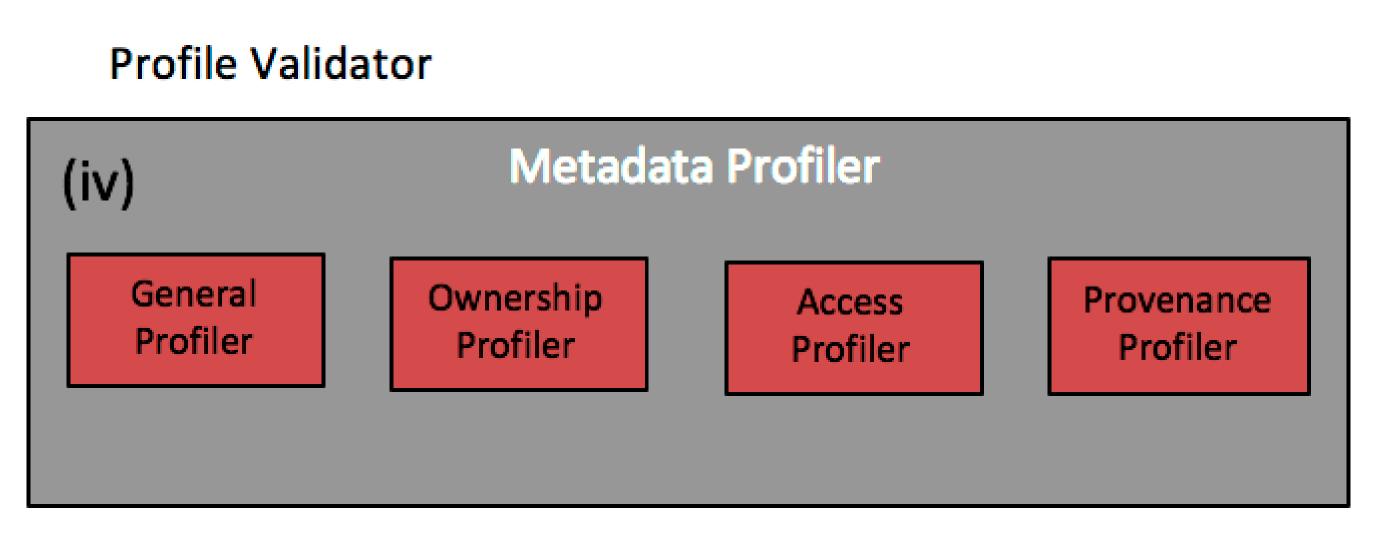


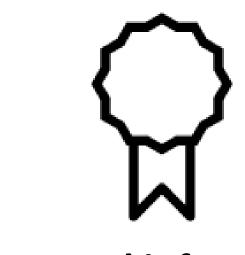




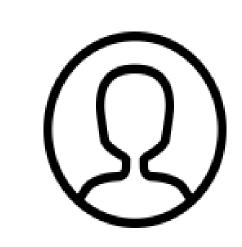








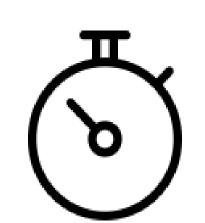
General information e.g. title, description



Ownership information e.g. author, maintainer_email



Access information e.g. license_title, license_id



Provenance information e.g. creation_date, version

Running Roomba on the LOD Cloud[†] group on the Datahub (259 Datasets)

41%

of the ownership information is missing

64%

of the resources metadata is missing 44%

of the top metadata problems fixed automatically

All results are reproducible from our code base at https://github.com/ahmadassaf/opendata-checker

This research has been partially funded by the European Union's 7th Framework Programme via the project Apps4EU (GA No. 325090)

