design_v2.tex

USAID) datasets

Data registration Module

Schematically store hosting URL, download URL, data dictionary, about URL, etc. of all datasets
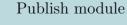
Data ingestion module

Crawler to fetch registered data (and metadata) into central data corpus. Finds storage format (CSV, XML, rasters, etc.) and syntactic entities (columns, cells, etc.)

Semantic annotation module

Semi-automatic means to extract metadata, social coordinates and indicators; annotate the discovered data using CRM concepts

Publish module

Query data bank; publish using APIs and web interface

Data bank

Structured a semi-structured data, and upload to the data bank

Validation & testing module

Compute and compare distributions; compute statistical measures and treands; compute outliars

Mapping & linking module

Semantic enrichment; further annotations and cross linking with hierarchies