Publishing the 15th Italian Population and Housing Census in Linked Open Data

Monica Scannapieco, R. M. Aracri, S. De Francisci,
A. Pagano, L. Tosco, L. Valentino

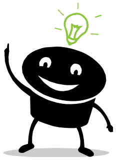
Istituto Nazionale di Statistica – Istat

NSIs & LOD

- NSIs produce data
  - Data dissemination is a fundamental phase
- Different models are adopted by NSIs to represent data:
  - DDI (Document Data Initiative) (1995)
  - Neuchâtel model (2004)
  - SDMX (Statistical Data and Metadata Exchange) (2004)
  - GSIM (Generic Statistical Information Model) (2013)

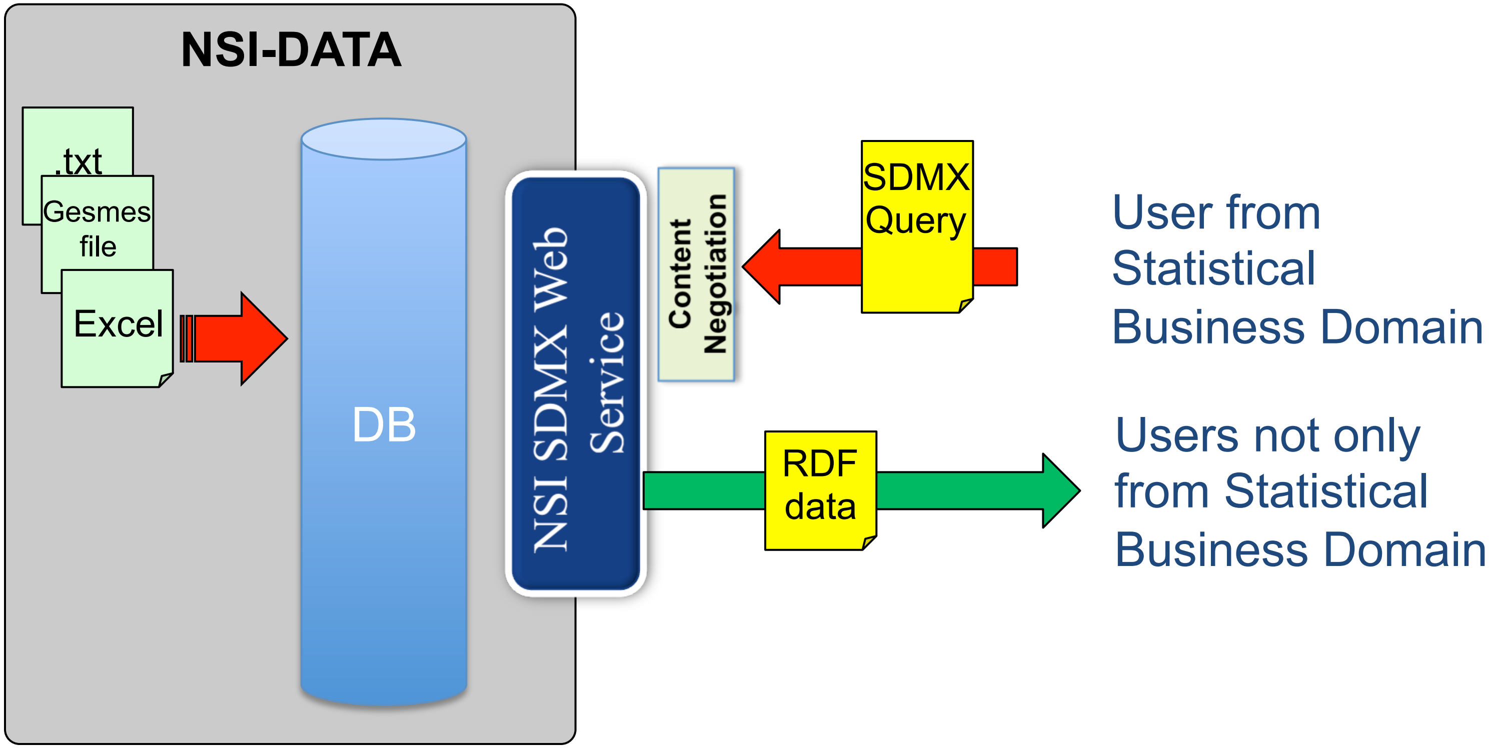
- Need to broaden the dissemination to non-statistical users

- Data dissemination in LOD format !

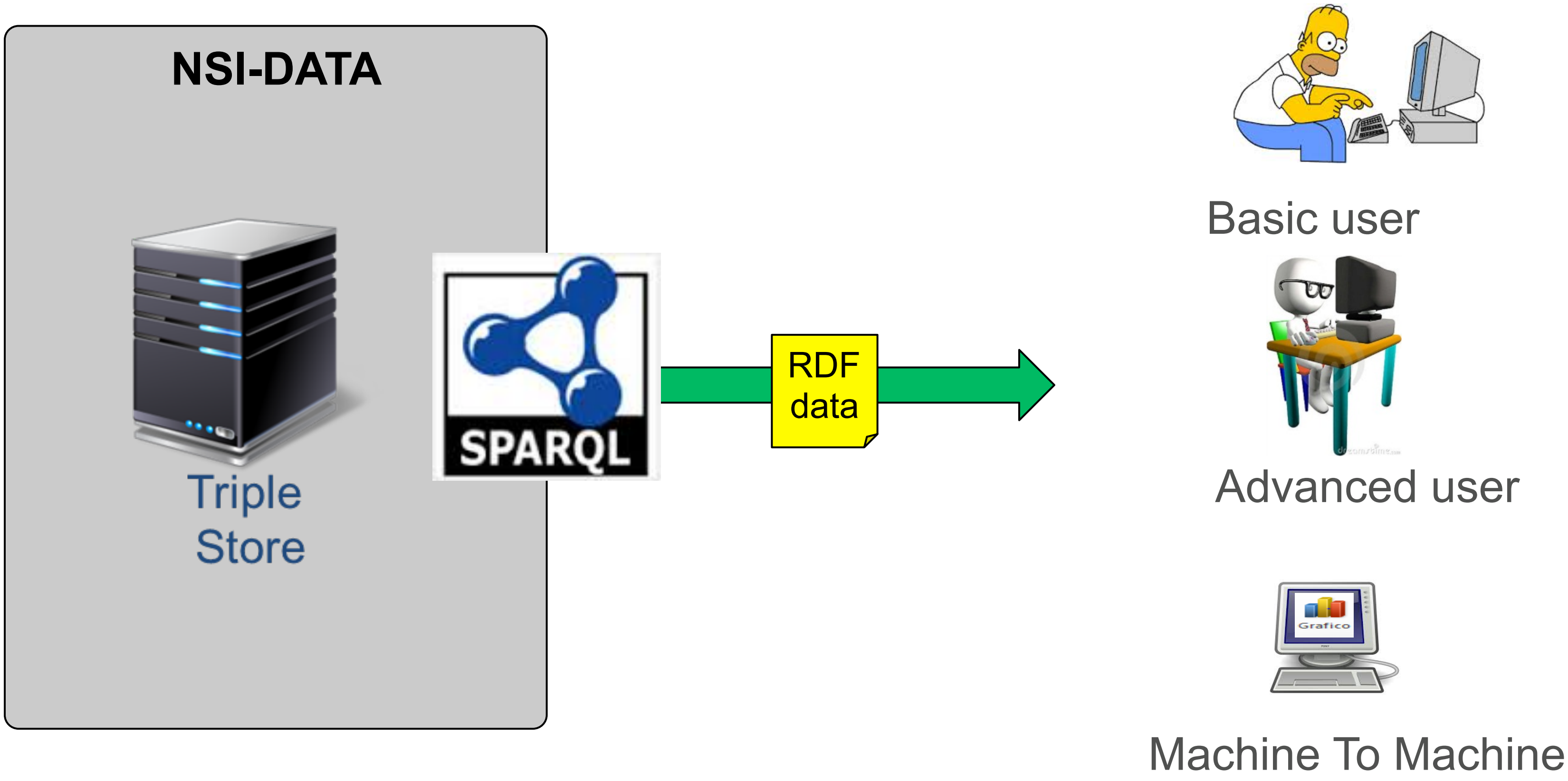


- Among NSIs who published data in LOD format:
  Istat (Italy), INSEE (France), ABS (Australia), EL.Stat (Greece), CSO (Ireland)

Istat SDMX-based Data Dissemination - 1



Istat LOD-based Dissemination - 2



NSI-DATA

Triple Store

RDF data

Basic user

Advanced user

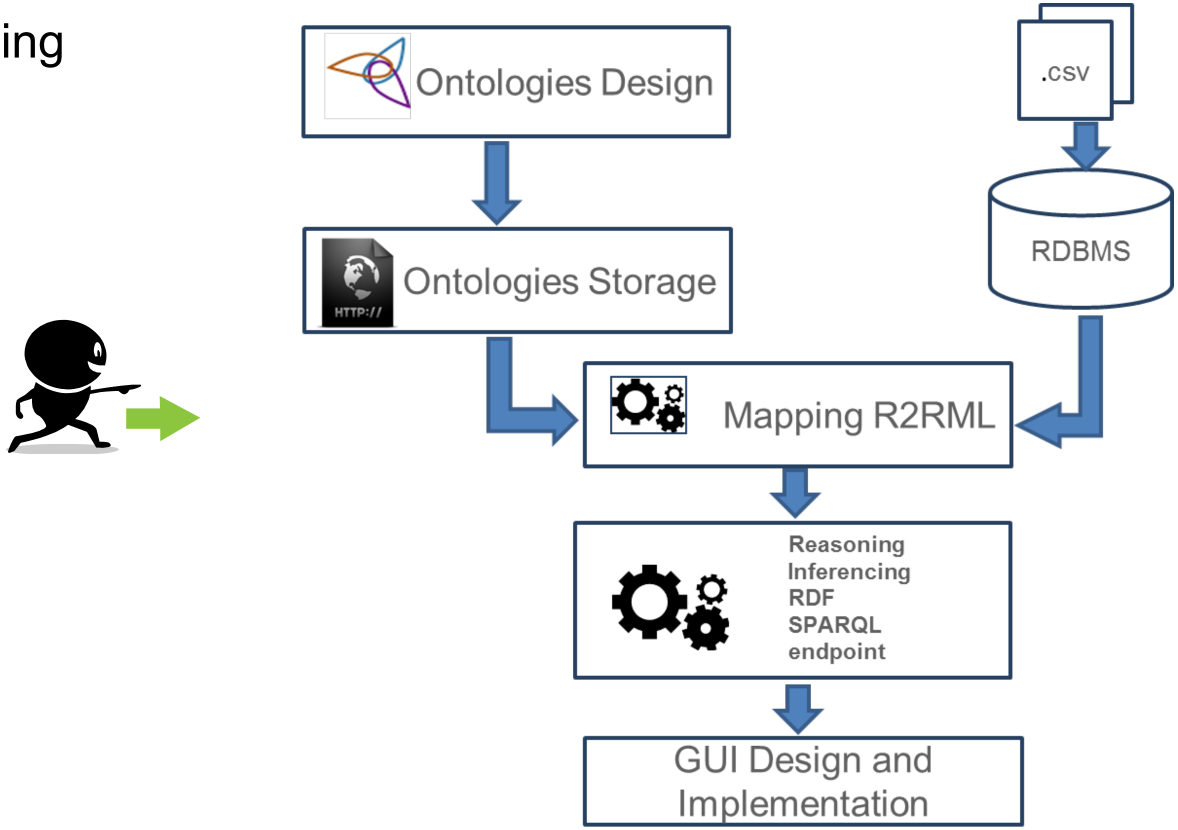
Machine To Machine

The CensLOD Project: Main Phases & Workflow

The project consists of three main phases:

1. Domain analysis and Ontology definition
2. Triples generation
3. LOD publishing

Project Workflow



Phase 1: Domain Analysis

- Territory dataset :describing the Italian territorial features from both administrative and geographical perspectives
- Censpop dataset: describing the population and housing Census indicators, at the territorial level of Census section
  - Published in the past as CSV files or as XLS files (http://www.istat.it/it/archivio/104317)

COD_PRO	COD_COM	PRO_COM	SEZ2001	SEZIONE	P1	P2	P3	P4	P5	P6	P7
5	1	5001	50010000005	5	9	6	3	3	4	0	2
5	5	5005	50050000343	343	34	17	17	12	15	2	5
5	118	5118	51180000013	13	13	7	6	5	5	1	1
5	120	5120	51200000001	1	292	141	151	104	133	7	45
5	121	5121	51210000037	37	23	11	12	10	8	0	4

Data size

- ✓ 402.903 Census Sections
- ✓ 74.482 Localities
- ✓ 2.200 Census Areas
- ✓ 3.631 Geomorphological entities
- ✓ And others classes …

- ✓ 43 indicators for each entity (currently loaded):
  - ✓ Resident Population – Males
  - ✓ Resident Population – age > 74 years
  - ✓ Foreigners and stateless persons resident in Italy – Males
  - ✓ …

Phase 1: Ontologies Definition

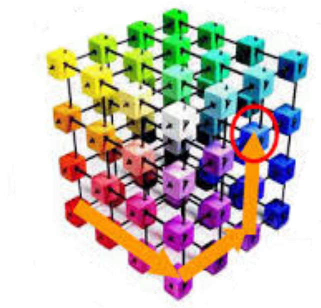
- Two distinct Ontologies:
    - Territorial Ontology
    - Census Data Ontology (population & housing)

- OWL Ontologies
- Use of Meta Ontologies:
    - ***SKOS and XKOS***: skos:Concept, …
    - **ADMS**: adms:AssetRepository, …
    - ***Data Cube Vocabulary***: qb:DataSet, qb:Observation, …
    - ***PROV***: prov:wasGeneratedBy, …
    - ***GeoNames:*** gn:name, gn:countryCode, gn:parentCountry, …

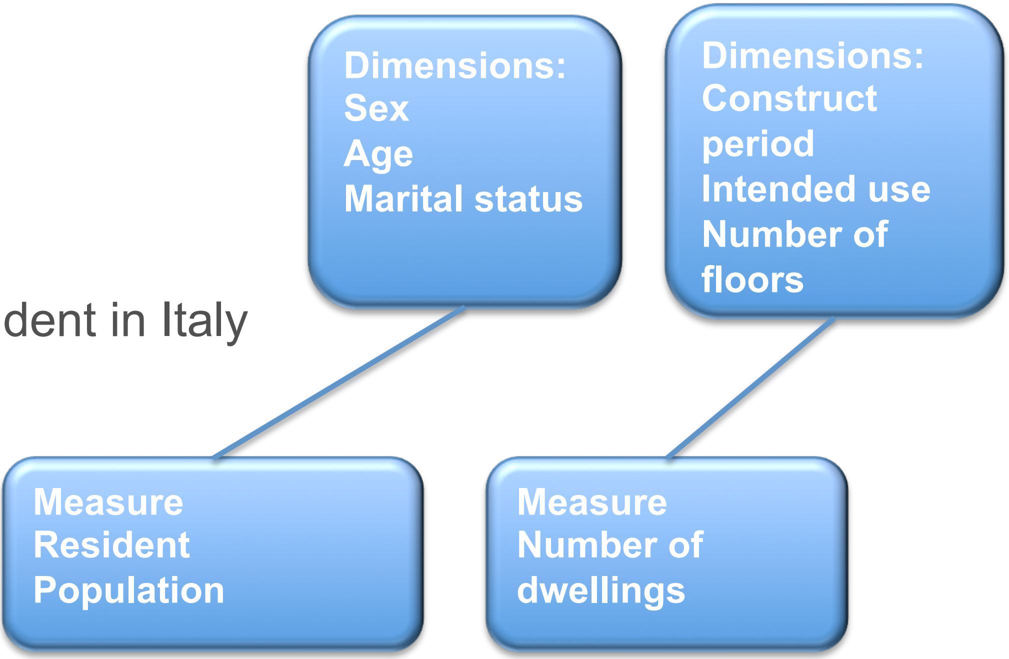
Territorial Ontology

- Description of principal classes of the domain:
  - 95 entities
    - Regions
    - Province
    - Locations
    - …..
  - 200 roles:
    - *appartiene ACDDASC* (links municipality with its sub-municipalities components)
    - *equivalentTo* (links *entity* with the relative Geonames entity)
    - …..

Census Data Ontology



- Use of RDF Data Cube Vocabulary that allows to publish multi-dimensional data
- Dimensions:
  - Sex
  - Age classes
  - Citizenship
  - Territory (territory defined in the Territorial Ontology)
  - Construction Period
  - Number of floors
  - …..
- Measures:
  - Number of residents
  - Foreigners and stateless resident in Italy
  - Number of Housing
  - …..



Dimensions:
Sex
Age
Marital status

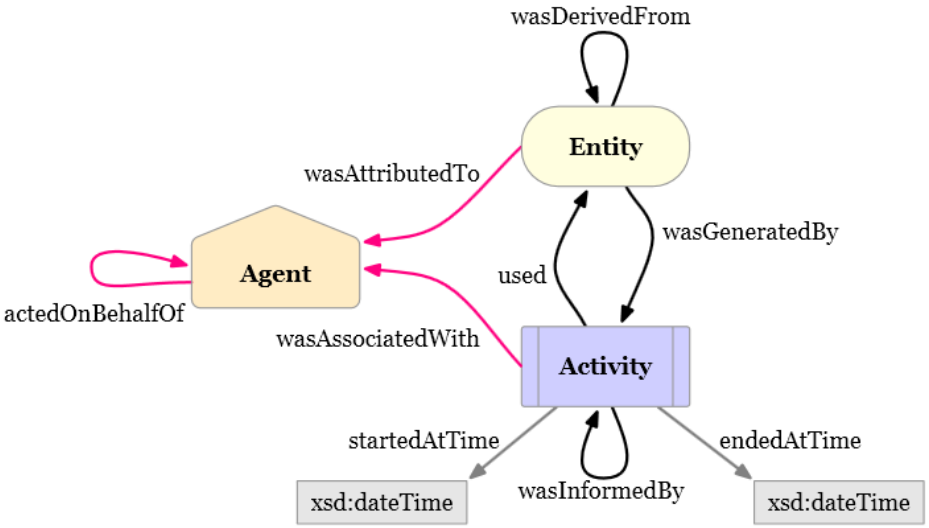
Dimensions:
Construct period
Intended use
Number of floors

Measure
Resident Population

Measure
Number of dwellings

Certifying Istat Data

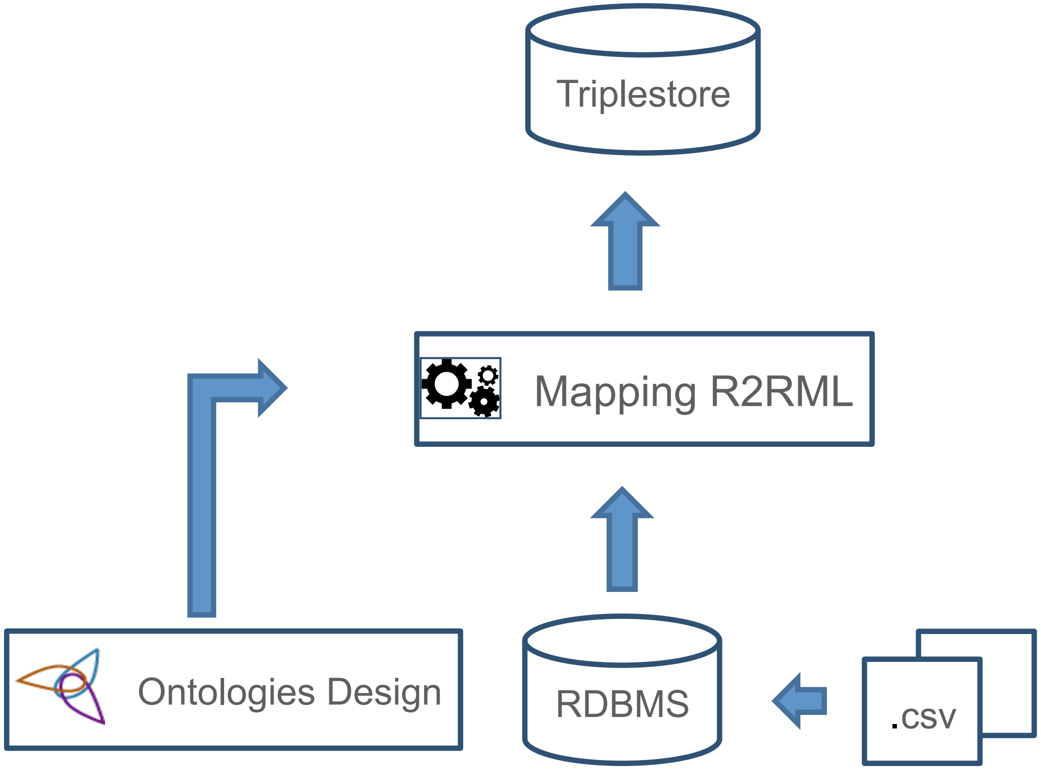
- Istat data are the results of established methodological procedures:
  - Official Statistics has a precise meaning in terms of quality and trust of the statistical information product



- We used the W3C PROV Ontology as a structured description of the provenance of the data
  - Where data come from
  - Official data sources according to European and National regulation
  - Domain standard conformance (e.g., variant and version of a statistical classification)
  - …

Phase 2: Triples Generation

- Triples generation by mapping CSV source file into triples using R2RML language (http://www.w3.org/TR/r2rml/)
  - Mapping rules manually written



Mapping Example

Example R2RML mapping

```
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix ex: <http://example.com/ns#>.
@prefix ter: <http://rdf.istat.it/ter/> .

<#TriplesMapZonaInContestazione>
   rr:logicalTable [ rr:tableName "ZONE_IN_CONTESTAZIONE" ];
   rr:subjectMap [
      rr:template "http://dati.istat.it/ter/ZonainContestazione/{COD_ZONA_C}";
      rr:class ter:ZonaInContestazione;
                        rr:class ter:AreaSpeciale;
   ];
   rr:predicateObjectMap [
      rr:predicate ter:contestatoDa;
      rr:objectMap [ rr:column "PRO_COM" ];
   ];

 rr:predicateObjectMap [
      rr:predicate ter:nomeAreaSpeciale;
      rr:objectMap [ rr:column “NOME_AS" ];
   ];
```

Mapping rules to obtain all the contested zones showing two properties:

- nomeAreaSpeciale
- contestatoDa

Result (Turtle)

```
<http://dati.istat.it/ter/ZonainContestazione/5>
     a       ter:ZonaInContestazione ,
ter:AreaSpeciale ;
     ter:contestatoDa "96001" , "2066" ;
     ter:nomeAreaSpeciale "Regione Folla" .
```

Mapping result

Contested zone #5, named ‘Regione Folla’, that is a AreaSpeciale and is contested between two municipalities identified by ‘96001’ and ‘2066’

Phase 2: Triples Generation

OWL2 Profiles syntactic subsets of OWL 2 - each is more restrictive than OWL DL

- OWL 2 EL
    - polynomial time algorithms
    - particularly suitable for applications
        - very large ontologies are needed
        - expressive power can be traded for performance guarantees
- OWL 2 QL
    - conjunctive queries to be answered in LogSpace (more precisely, AC0) using standard relational database technology
    - particularly suitable for applications
        - relatively lightweight ontologies are used to organize large numbers of individuals
        - it is useful or necessary to access the data directly via relational queries (e.g., SQL)
- OWL 2 RL
    - implementation of polynomial time reasoning algorithms using rule-extended database technologies operating directly on RDF triples
    - particularly suitable for applications
        - relatively lightweight ontologies are used to organize large numbers of individuals
        - it is useful or necessary to operate directly on data in the form of RDF triples

Phase 2: Triples Generation

- **Usage of *Oracle Spatial and Graph***
  - Vocabularies & rulebase (for inferencing) supported

OWL with IF Semantic [1]

1. RDF++
-all RDFS vocabulary constructs
-owl:InverseFunctionalProperty
-owl:sameAs

2. OWLSIF
-all RDFS vocabulary constructs
-owl:InverseFunctionalProperty
-owl:TransitiveProperty
-owl:inverseOf
-owl:equivalentProperty
-owl:someValuesFrom
-owl:FunctionalProperty
-owl:SymmetricProperty
-owl:sameAs
-owl:equivalentClass
-owl:hasValue
-owl:allValuesFrom

3. OWL Prime
-rdfs:subClassOf
-rdfs:domain
-owl:FunctionalProperty
-owl:SymmetricProperty
-owl:sameAs owl:inverseOf
-owl:equivalentClass
-owl:hasValue
-owl:allValuesFrom
-owl:disjointWith
-rdfs:subPropertyOf
-rdfs:range
-owl:InverseFunctionalProperty
-owl:TransitiveProperty
-owl:equivalentProperty
-owl:someValuesFrom
-owl:differentFrom
-owl:complementOf

Supported semantics for these value restrictions are only intensional (IF semantics).

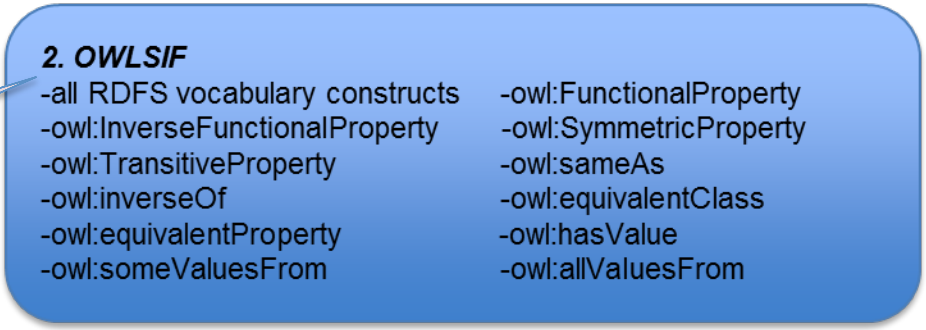
4. OWL 2RL & 5. OWL2 EL
As described in
http://www.w3.org/TR/owl2-profiles/#OWL_2_RL
http://www.w3.org/TR/owl2-profiles/#OWL_2_EL

[1] *Completeness, decidability and complexity of entailment for RDF Schema and a semantic extension involving the OWL vocabulary*, by H.J. Horst, Journal of Web Semantics 3, 2 (2005), 79–115

Phase 2: Triples Generation

- **Inference rulebase chosen**
  - OWLSIF

OWL with IF Semantic



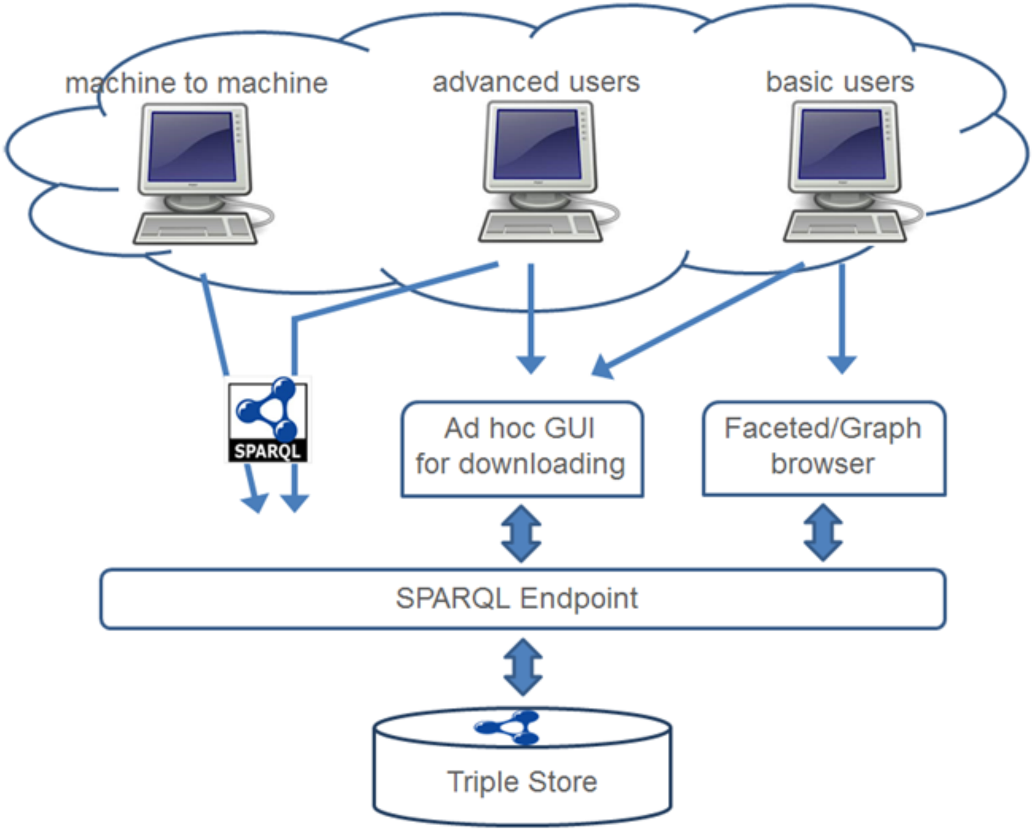
2. OWLSIF

-all RDFS vocabulary constructs
-owl:InverseFunctionalProperty
-owl:TransitiveProperty
-owl:inverseOf
-owl:equivalentProperty
-owl:someValuesFrom
-owl:FunctionalProperty
-owl:SymmetricProperty
-owl:sameAs
-owl:equivalentClass
-owl:hasValue
-owl:allValuesFrom

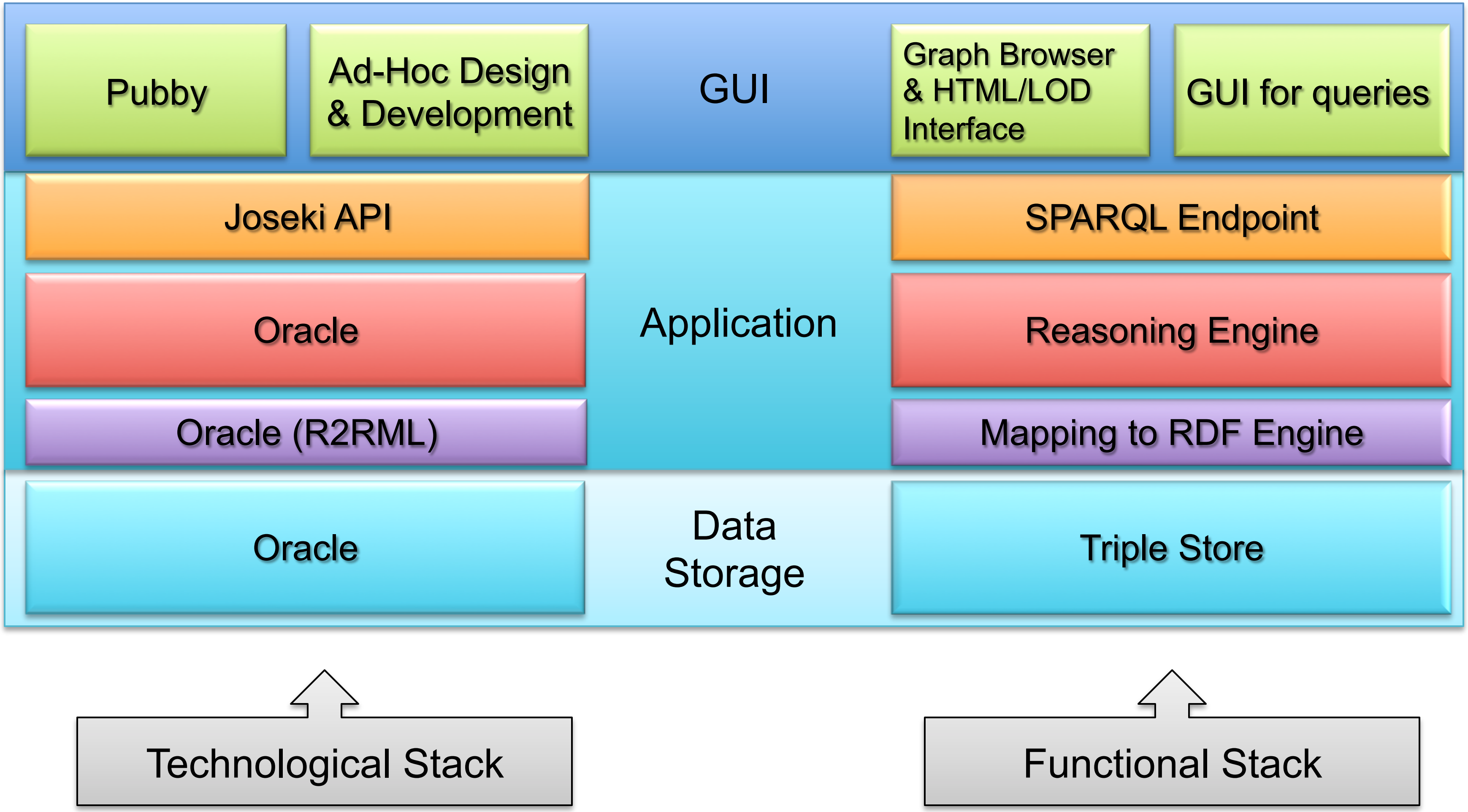
- **Optimizations applied**
  - Triples materialization
    - Considerable reduction of response time
  - Definition of *"ad hoc"* inference rules
  - Storage of frequent queries

Phase 3: LOD Publishing

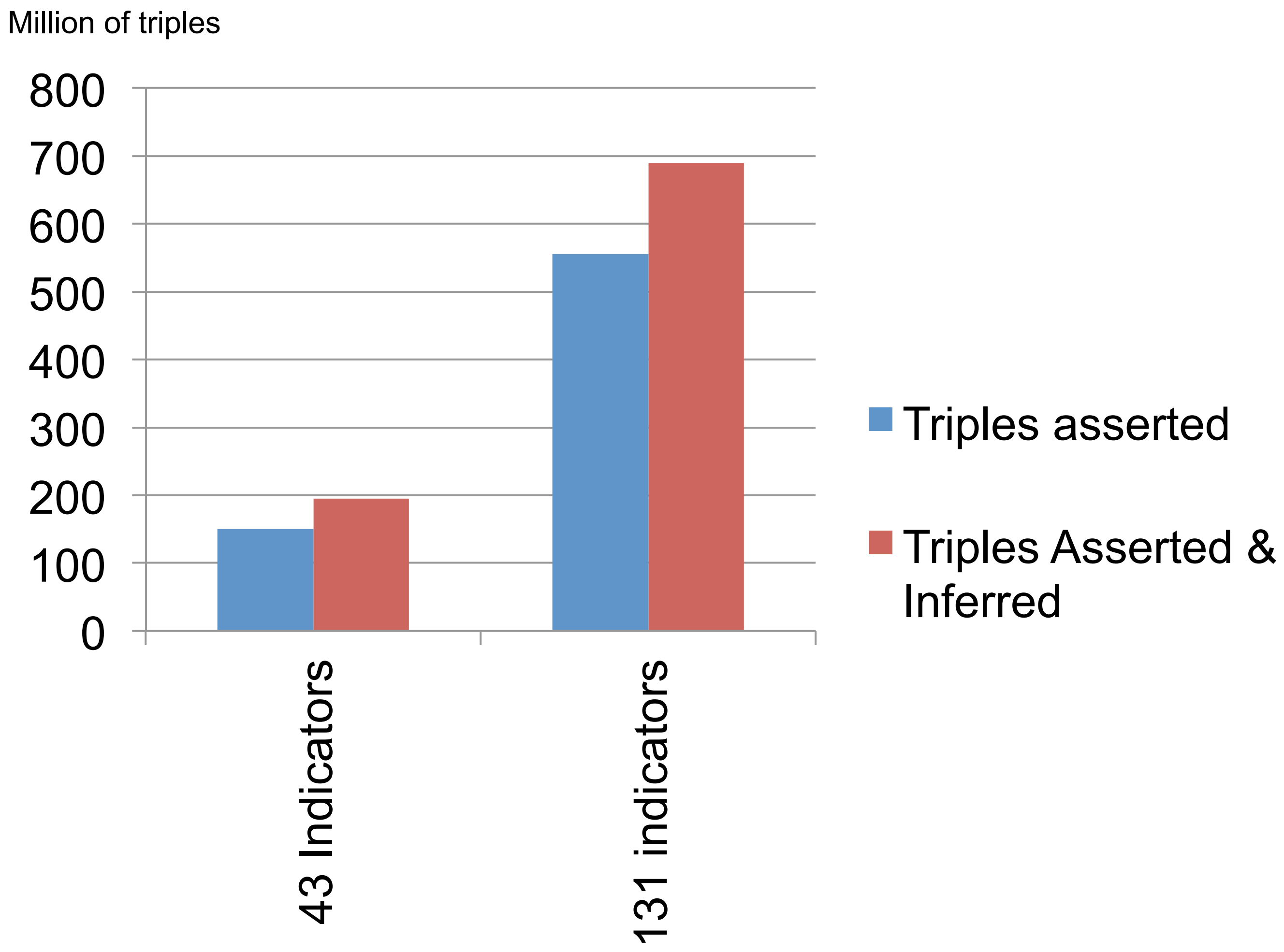
- Three access points to cover the requirements of the different possible users:
  - SPARQL endpoint
    - Advanced users
    - Machine-to-machine communications
  - Linked Data Interface (Faceted/Graph browser)
    - Basic users
  - Ad-hoc GUI for datasets downloading
    - Basic users



Technological Environment



Performances



Concluding Remarks

- Cens-LOD is the first production process that deploys Istat data on an Istat SPARQL Endpoint
  - 2014: Publication of CensPop and Territory
  - 2015: Addresses

- LOD-based data dissemination will allow:
  - Machine-to-machine data provisioning by Istat (currently only SDMX datasets)
  - Widening the range of Istat data users
  - Improving efficiency of data exchange flows with Italian administrations

- …and much more ☺: like having the knowledge «Eaten»

By Mark Johnstone