IPHES, URV, Tarragona, Enero 2014

Closing the Knowledge Loop

V Workshop en Econosociofísica

Jesús Marco de Lucas

Profesor de Investigación del CSIC

Instituto de Física de Cantabria (IFCA)

CENTRO MIXTO CSIC-UNIVERSIDAD DE CANTABRIA

marco@ifca.unican.es

Background: Research at IFCA

- IFCA is a Research Center in Cantabria, in Santander (Campus UC)
  - CSIC – University of Cantabria
  - Basic Research:
    - Astrophysics, Particle Physics, Dynamics and fluctuations in nonlinear systems, Meteorology
- Research line on Advanced Computing and e-Science
  - E-Infrastructure support to large projects
    - EUROPEAN GRID / NATIONAL GRID
    - TIER-2 CMS / LHC
    - ALTAMIRA NODE IN SPANISH SUPERCOMPUTING NETWORK
      - 150 nodes, IB
      - 1 Petabyte storage & Archiving
      - INTEL x86, POWER7, GPU, IB, LARGE RAM
    - CLOUD ENABLED RESOURCES
      - FEDERATED CLOUD AT EUROPEAN LEVEL
  - ~~INTER~~MULTIDISCIPLINARY PROJECTS
    - CONUS
    - INTEGRAL TRACEBILITY

- **We are open to support (for "free") any good project with computing needs:**
  - **Through EU GRID or through RES or through new CLOUD pilot initiatives**
  - **DIRECTLY**

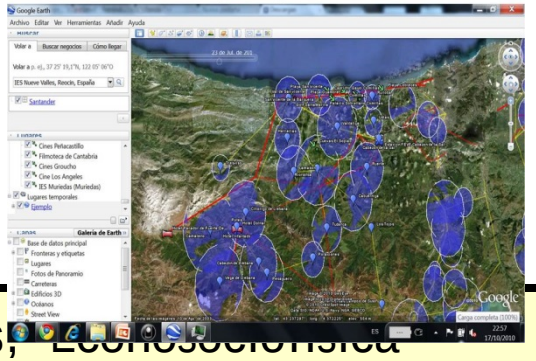
A first experience (2009-2010)

- An innovative SME company (CIC SL in the Parque Científico-Tecnológico de Cantabria) proposed to explore a "realistic" model to track people in the region using the information from mobile phones positioning (from cell towers)
  - Program to "track" up to 1M people with a time step of 5 minutes
  - Technical key: scalable (distributed) database (MongoDB)
  - Simulation: focus on tourism, using a "realistic model" of the region: Cities/Villages, Roads, Hotels, Restaurants, Beaches, Museums…
- We built a nice (although simplified) model, but did not validate it!
  - No way to get data from mobile phones!
- DESIGN AND VALIDATION require REAL INFORMATION
  - Even "trivial" information as where a hotel is placed, and its capacity

Sistema escalable

- Base de datos de tickets MongoDB
  - Hasta 4 instancias / nodo (multicore, large RAM)
  - Múltiples nodos balanceados
- Potencia de procesado distribuido
  - Uso de nodos en entorno GRID
  - **Simulación High Throughput**
    - **Ej: 100 nodos simulando actividad 1000 usuarios cada uno**
  - Postprocesado de queries complejas
  - **Agregación de información**

Output gráfico



2010-2013

- DATA PRESERVATION IN CMS (LHC EXPERIMENT)
  - How to guarantee that LHC data will be “available” by 2030 (as OPEN DATA)
- LIFEWATCH (MINECO, National LW Initiatives, CSIC)
  - EUROPEAN ESFRI, VIRTUAL LABS FOR BIODIVERSITY
  - DATA CYCLE AND REFERENCE MODEL
  - INTERDISCIPLINARY: FROM GENE TO SPECIES TO ECOSYSTEM
  - BIG DATA TOPICS (integration of heterogeneous info, real time)
  - ORIENTED TO PUBLIC MANAGEMENT OF NATURAL RESOURCES
- DORII (FP7) + ROEM+ (LIFE) with Ecohydros (SME)
  - Integration of remote data
  - Full Modelling
- An agent model to simulate anthropogenic impact
  - Farming
  - Village activity

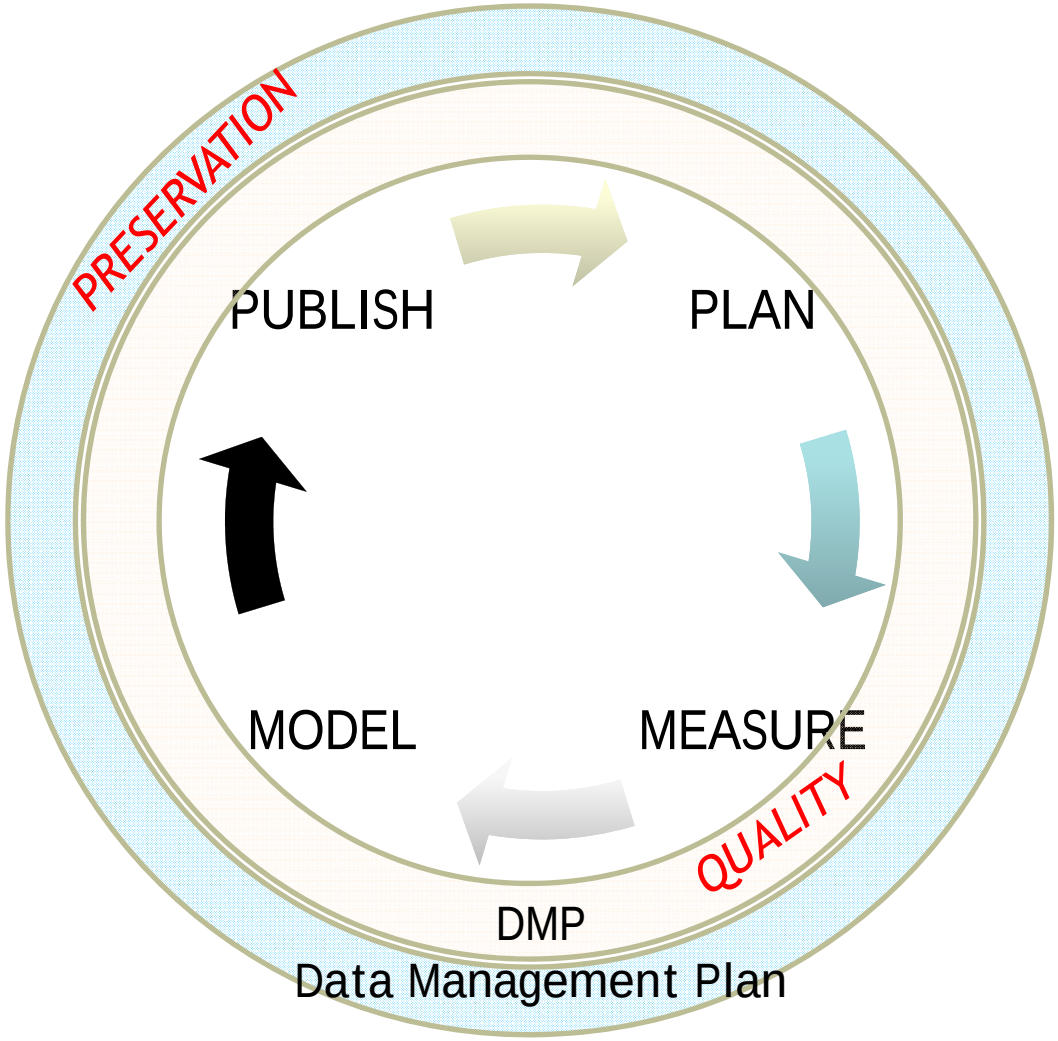
2014

- Abstract for this Workshop:

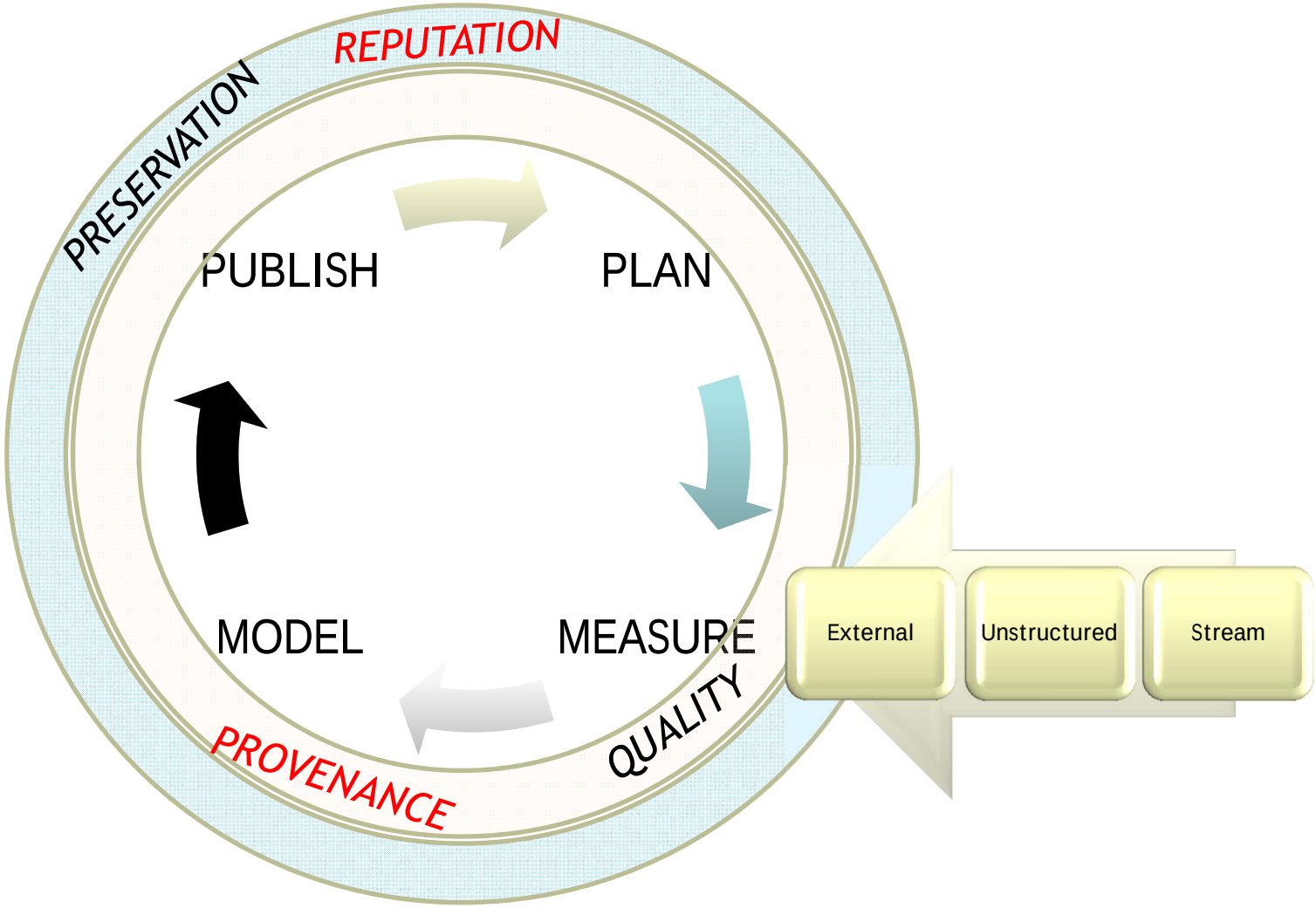
CLOSING THE KNOWLEDGE LOOP

Big Data infrastructures and techniques can be used to populate knowledgebases, and also to support more realistic agents based models. We discuss how to integrate both components to define a knowledge use & preservation framework.

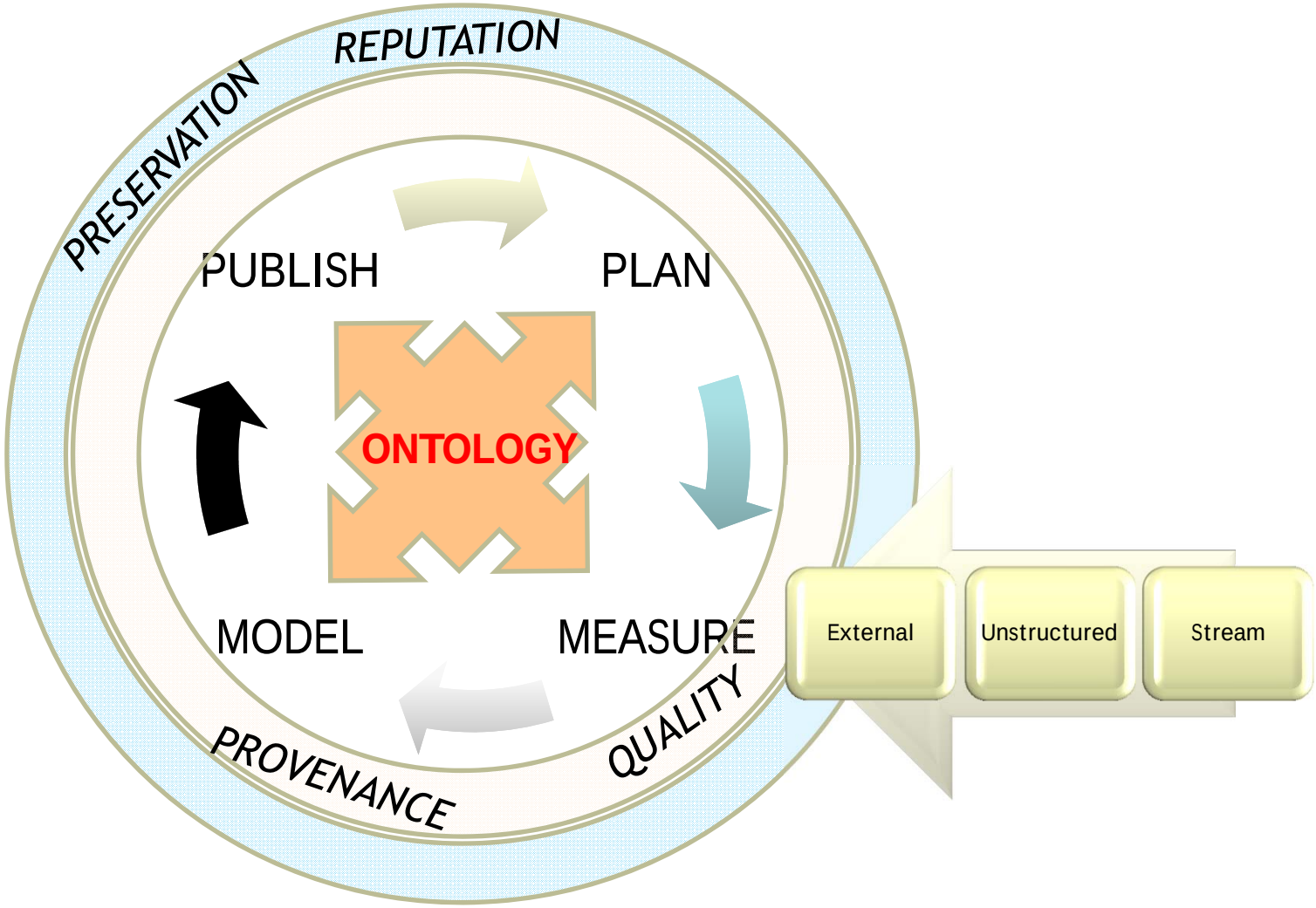
Knowledge Loop



Extended Knowledge Loop

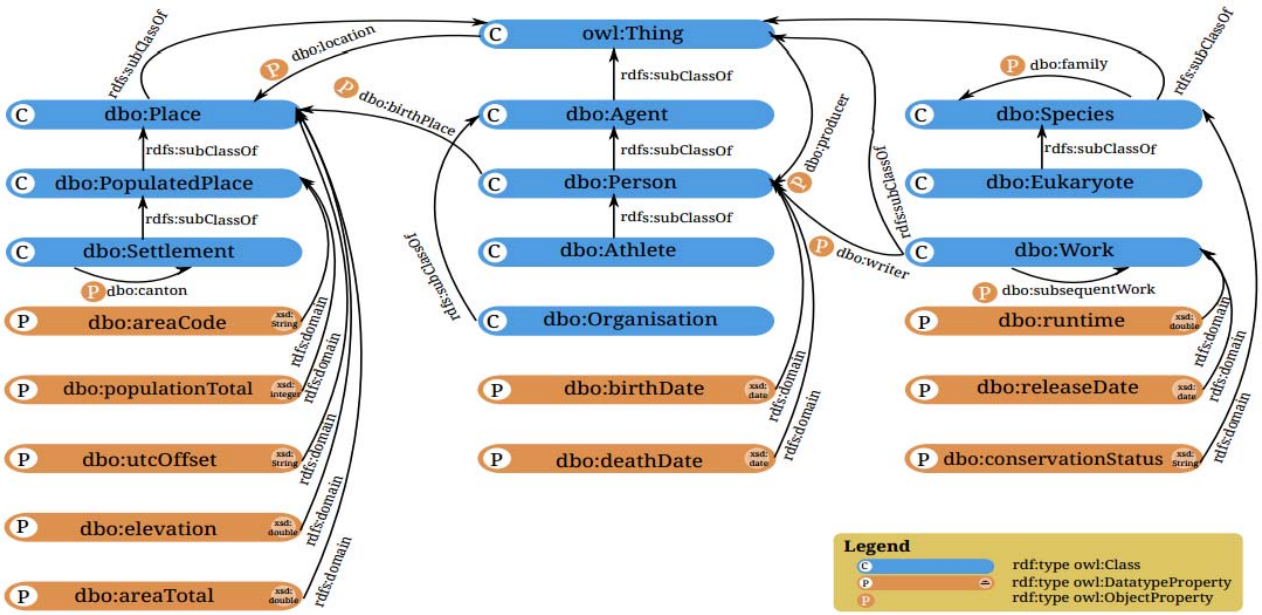


Closing the Knowledge Loop



Ontologies

In computer science and information science, an ontology formally represents knowledge as a set of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts



Snapshot of a part of the DBpedia ontology, from Semantic Web 1 (2012)

Knowledge & Ontologies

- It is a fact that (human) KNOWLEDGE is now lost in the loop
  - Due to the unavoidable/desirable evolution of Research Team/Methods
- COMMON VOCABULARY is key to inter/multi-disciplinary research
  - But also a source of fragmentation
  - BROKERS/TRANSLATORS are possible (as for DATA/METADATA)
- PLAN, MEASURE, MODEL & PUBLISH USING A (COMMON) ONTOLOGY?
- CAN WE INCLUDE TRUST/PROVENANCE?
- DOES AN ONTOLOGY DEFINE ADEQUATE AGENTS?
- WHAT IF MEASURE INCLUDES BIG DATA?
  - Exploit the idea of KNOWLEDGE BASES
  - Extend with trust/provenance and reputation
  - Example: from WIKIPEDIA to DBPEDIA
    - User sees RDF + SPARQL, cf. YAGO
    - (Now also as "TABLES")

- DO WE HAVE RESOURCES TO CREATE KB? YES

How to build a KB...

Lehmann et al. / DBpedia

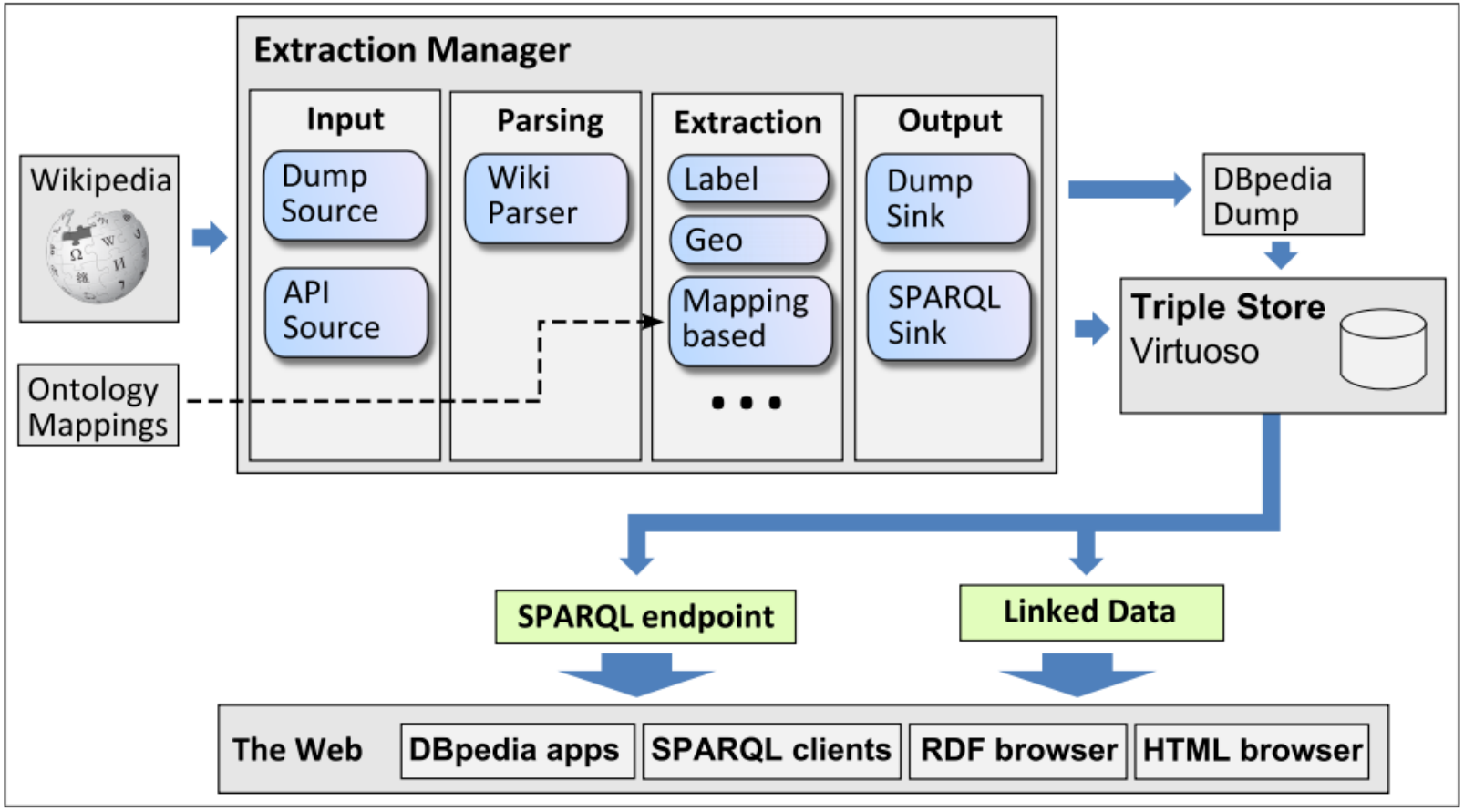


Fig. 1. Overview of DBpedia extraction framework.

A Challenge

- THE PROPOSAL:
  - **IDENTIFY POTENTIAL SOURCES FOR TRUSTED KNOWLEDGE BASES**
  - **BUILD KNOWLEDGE BASES**
  - **DEPLOY AGENTS MODELS TO EXPLOIT KNOWLEDGE BASES**

- Ejemplo:
  - 29 Enero 2014: 24.6% de la Economía en España es “sumergida”
    - Extremadura +++, Madrid ---
  - Solución propuesta: x3 número de inspectores de Hacienda
  - Es posible (para Hacienda) contar con un modelo basado en su propio BIG DATA que permita estimar las fuentes de economía sumergida???
    - CONTAR CON UNA ONTOLOGIA COMUN QUE DEFINA A UN CIUDADANO
      - Trabaja, Declara, Consume
    - CONSTRUIR KNOWLEDGE BASE ECONOMICA USANDO INFORMACION A TODOS LOS NIVELES / FUENTES
    - TESTEAR MEDIANTE UN MODELO DE AGENTES