SICRaS: a semantic big data platform for fighting tax evasion and supporting social policy making

Giovanni Adinolfi, giovanni.adinolfi@eng.it, Engineering Tributi SPA, Via G.B. Trener 8, 38121 Trento, Italy

Paolo Bouquet, bouquet@okkam.it, Okkam SRL, Via Segantini 23, 38121 Trento, Italy

Stefano Bortoli, bortoli@okkam.it, Okkam SRL, Via Segantini 23, 38121, Trento, Italy

Lorenzo Zeni, lorenzo.zeni@alysso.it, Alysso SRL, Via G.B. Trener 8, 38121 Trento, Italy

Introduction: the business needs

In these years, Italy is dealing with serious economic and social issues, which are aggravated by the recent global economic crisis. These issues include, among others: high fiscal burden, widespread tax evasion, increasing unemployment rate and the progressively aging of population. In this context there are two major needs that public administrations have to meet. On one hand, it is mandatory to fight tax evasion, on the other hand, there is the need to ensure social services in a more efficient, fair and effective way.

Our industrial need is to support governance and policy making in achieving these goals through the integrated analysis of a large amount of information collected by both public administrations and other public officers and organizations (e.g. notaries, public utilities and so on). This information is typically scattered over several heterogeneous and decoupled data sources. Moreover, it might also be partially outdated, unreliable and redundant.

The adoption of semantic technologies enables the construction of an integrated, trustworthy and accurate knowledge base, providing a picture of the fiscal and social situation of each single citizen and of the community of a local municipality, overcoming the limitations of legacy systems. Cornerstones of the solution are: 1) a set of domain ontologies aimed at improving data integration, and at producing useful inference from explicit information; 2) a scalable system to reconcile identities to the same real world entity across datasets, associating a unique and persistent name to each single entity; and 3) leveraging geo-spatial technologies to achieve a deeper understanding of the observed districts by means of spatial analysis and reasoning.

Towards a Linked Data for tax domain

Tax information systems typically work with an extraordinary amount of data concerning many different aspects of taxpayer's life: personal and company details, cadastral information, job positions and so on. The role of ontologies in data exchange and integration so as to retrace tax positions from all these information is definitely invaluable [1]. In fact, since data are collected at various times by different partner administrations, there is the need to link all these tax related information from distributed source streams. Looking at the domain, we recognize that tributes are generally imposed taking into account of specific circumstances or events that happen during the taxpayer's life: the taking up of residency in a new town, join a nuclear family, buying a house, etc.

Given these assumptions, we chose to follow an entity/event based ontological modelling approach. The goal is to support an integration pipeline, producing a unified view of the relevant fiscal facts scattered in datasets supplied by diverse public institutions. The entity/event based modelling approach allows then to materialize, at a given instant of time, the deduced tax position involving each single taxpayer.

The Pipeline from Data to (Big) Knowledge

One of the known limits of semantic technologies is scalability both in reasoning and data management. This has twofold justification, on the one hand there are limits related to the computational complexity of reasoning based on model theoretic semantics, on the other hand the immaturity of existing technologies for data management. To

overcome these limits, we organized a pipeline relying on an ensemble of scalable and state of the art technologies to define a Semantic ETL suitable to create a Semantic Big Data Pool.

We rely on a customized and optimized version of Open Refine tool to perform data cleaning operation, including syntactical validations and transformation of the original data coming from the public institutions. The formal and syntactical validations, expressed according to a specific rule language, are the results of many years of experience on the field. At this stage, issues related to semantic and structural heterogeneity affecting the original data are normalized relying on a set of maintainable contextual ontology mappings towards the defined domain ontology. Each record is analysed to extract information about the involved entities to reconcile their identities relying on the Okkam Entity Name System [2]. Once the identity of any entity involved in each of the records has been disambiguated, the dataset is exported in RDF and stored as many entity-centric named graphs into the Hadoop Distributed File System (HDFS).

The result of this first part of the process is a physically distributed and logically integrated large RDF graph that can be manipulated and processed relying on emerging big data technology. Therefore, we rely on tools like Apache HBase, Apache Incubated Spark, Apache Incubated Flink, Apache Hive, and Apache Pig to define (complex) big data shuffling processes producing any view, analysis and mesh-up necessary to support tax assessment domain applications. In fact, it is possible to select subsets of the giant RDF graph to store it in application specific data management systems. For example, we sink data into a triple store such as OpenRDF Sesame and enable scalable rule-based reasoning tasks using SPRINGLES. Another example is to build sub-graphs to support seamless real time navigation of the knowledge relying on effective indexing tools such as as Apache SIREn, or to perform graph-based analysis sinking data in a graph database (e.g. Neo Technology Neo4J). Finally, it is possible to integrate semantic technologies in the core of the SpagoBI suite in order to enable novel Business Intelligence tools and techniques [3].

Geographic technologies for spatial analysis and reasoning

The integration of geo-spatial technologies adds an important analytic dimension to SICRaS. Taking advantage of this kind of information, we firstly intend to exploit the notion of territory, seen as a spatial region in our ontological model. This enables new ways of extracting, observing and analysing data about real world entities and the spatial relations among them. Secondly, we develop techniques to match entities relying on geo-spatial features to link our knowledge base to external sources (e.g. urban development plans) to find out new information valuable for tax assessment and for other fiscal and social purposes.

Concluding remarks

In SICRaS we define a scalable and efficient data processing pipeline, capable of overcoming the limits of current semantic technologies riding the wave of emerging big data processing tools. A wise union of semantic and big data technologies, tempered with deep domain knowledge and sophisticated geospatial tools, creates seamless opportunities to define tax assessment applications. Exploiting the wealth of data in an efficient and effective way, we aim to define the next generation of tools for policy makers and help the Italian institutions in overcoming the challenges of the 21st century.

References

[1] Isabella Distinto, Nicola Guarino, and Claudio Masolo. 2013. A well-founded ontological framework for modeling personal income tax. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law* (ICAIL '13). ACM, New York, NY, USA, 33-42.

[2] Paolo Bouquet, Heiko Stoermer, Claudia Niederee, and Antonio Maña . 2008. Entity Name System: The Back-Bone of an Open and Scalable Web of Data. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing* (ICSC '08). IEEE Computer Society, Washington, DC, USA, 554-561.

[3] Matteo Golfarelli. 2009. Open Source BI Platforms: A Functional and Architectural Comparison. In *Proceedings of the 2009 International Conference Data Warehousing and Knowledge Discovery (DaWaK 2009) Linz, Austria, August 31–September 2, 2009* . Springer Berlin Heidelberg, 2009, 5691, 287-297