ODINet - Online Data Integration Network

An innovative ontology-based data search engine

S. Pieroni, M. Franchini, S. Molinaro
Institute of Clinical Physiology, CNR
Pisa, Italy
{s.pieroni, m.franchini, molinaro}@ifc.cnr.it

A. Greco, F. Pitto
Sistemi Territoriali S.r.l.
Cascina (Pisa), Italy
{a.greco, f.pitto}@sister.it

M. Toigo
Simurg Ricerche snc
Livorno, Italy
m.toigo@simurgricerche.it

L. Caterino
Rete Sviluppo S.C.
Firenze, Italy
caterino@retesviluppo.it

Abstract— Along with the expansion of Open Data and according to the latest EU directives for open access, the attention of public administration, research bodies and business is on web publishing of data in open format. However, a specialized search engine on the datasets, with similar role to that of Google for web pages, is not yet widespread. This article presents the Online Data Integration Network (ODINet) project, which aims to define a new technological framework for access to and online dissemination of structured and heterogeneous data through innovative methods of cataloging, searching and display of data on the web. In this article, we focus on the semantic component of our platform, emphasizing how we built and used ontologies. We further describe the Social Network Analysis (SNA) techniques we exploited to analyze it and to retrieve the required information. The testing phase of the project, that is still in progress, has already demonstrated the validity of the ODINet approach.

Keywords-Data search engine; domain ontology; semantic web; social network analysis.

I. INTRODUCTION

The Online Data Integration Network (ODINet) project is a Research and Development Project, approved within the Italian Regional Operational Programme having as the main objective the "Regional Competitiveness and Employment" through the 2007-2013 European Regional Development Fund. The project involves the prototypal implementation of an innovative semantic search engine (SSE) able to catalog numerical data in an ontological graph, to extract from data the information most relevant to the user requests using Social Network Analysis (SNA) algorithms and to return that information in a highly usable way.

The literature review indicates that there are multiple proposals for SSE, but none of them is specialized in finding information contained in alphanumeric datasets related to the user's needs. On the other hand, there are some search engines for numerical datasets, such as Quandl (https://www.quandl.com) and datahub.io (http://datahub.io), but it seems they are neither based on a semantic knowledge base, nor they employ SNA techniques.

Therefore, the ODINet's design is based on these prerequisites and its goal is to demonstrate the benefits of the combined use of these innovative principles.

The application domain is connected to the social, economic and health fields, in order to cover most of the available data held by public bodies in the Italian context. Moreover, the three domains are closely linked to one another and offer the opportunity of cross-sectional cognitive investigations, through the identification of ontologies that describe interconnected concepts and links among various topics. Within each area, a kernel ontology has been designed for automatic and manual indexing purposes, bringing direct support to the SSE and enhancing retrieval accuracy. We further developed a data harvest component able to extract data from the web, interfacing to open data portals of Italian public administration.

We indexed those datasets together with the thematic ontologies, building a Search Graph that constitutes the main support for the Search Engine. This component, that is available as a web service, exploits well-known algorithms based on SNA properties, such as the centrality of nodes and the clusterization factor, in order to identify those datasets that are more related to the search query inserted by the user.

The search engine performs a semantic search on the graph: the semantic relations of our ontologies are enriched with two further procedures able to identify concepts in distinct ontologies that are semantically related one another. Finally, the identified datasets semantically connected to the user's query are returned by the web interface. A diagram describing the overall organization of ODINet platform is shown in Figure 1.

The focus of this paper is primarily the description of the semantic component of our platform, emphasizing the ontology-building process and how the ontologies have been used to build an index in form of a graph. We further present an in-depth description of our search engine component, explaining the whole process from a search query inserted by a user to the final results returned by the search engine.

The rest of this paper is organized as follows. Section II describes the knowledge resources and the ontology building process. Section III describes the technological platform and Section IV shows the main results. A final discussion closes the article.

II. KNOWLEDGE RESOURCES AND ONTOLOGY BUILDING

Building specialized ontologies from scratch through the support of domain experts' knowledge, requires a huge effort of conceptualization and a long editing time [1], especially in complex domains. Therefore, our choice was to rely on existing standard resources starting from EuroVoc [2], a

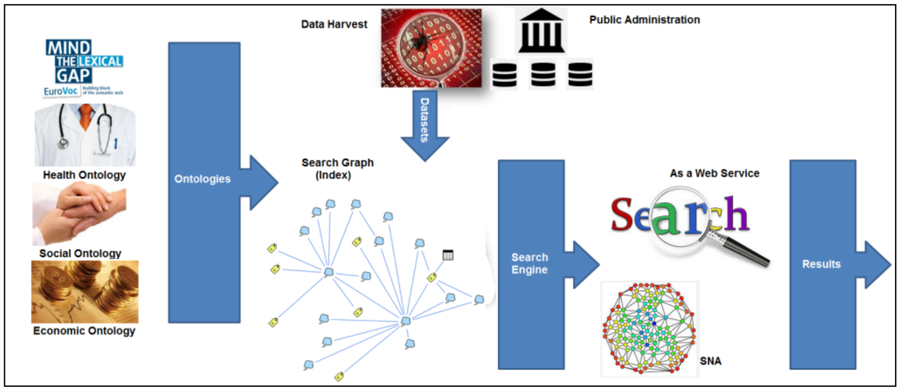
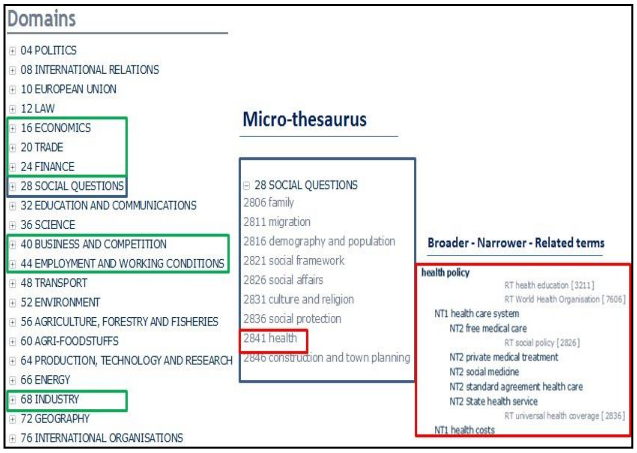


Figure 1. Overall organization of the technological platform

multilingual, multidisciplinary thesaurus, managed and maintained by the European Union’s Publications Office, which moved forward to ontology-based thesaurus management and semantic web technologies as Simple Knowledge Organization System (SKOS) conformant to W3C recommendations. EuroVoc has been widely used for classification software and indexers development [3], therefore starting from the SKOS [4] version of EuroVoc and relying on other standard resources, three specialized ontologies have been developed and linked in order to support the SSE. Since the ODINet's main objective is to access and classify a large amount of data and to present it to a wide range of users, EuroVoc has been chosen as it covers an exhaustive set of fields related to the activities of the European institutions, as shown in Figure 2. Therefore, in the initial phase of our work, the ontologies have been projected to have a wide horizontal spectrum, rather than a vertical one, using a top-down approach aimed to develop precise definitions of high-level concepts, and postponing to the validation phase a bottom-up approach necessary for analytic use-cases. After a deep analysis of the various sectors, we have chosen the ‘Social Questions’ domain as the core resource for both social and health ontologies. Actually, this domain copes with various relevant topics for the ODINet project as health, family, migration, demography, social framework and affairs, culture and religion and social protection.



Domains

⊞ 04 POLITICS
⊞ 08 INTERNATIONAL RELATIONS
⊞ 10 EUROPEAN UNION
⊞ 12 LAW
⊞ 16 ECONOMICS
⊞ 20 TRADE
⊞ 24 FINANCE
⊞ 28 SOCIAL QUESTIONS
⊞ 32 EDUCATION AND COMMUNICATIONS
⊞ 36 SCIENCE
⊞ 40 BUSINESS AND COMPETITION
⊞ 44 EMPLOYMENT AND WORKING CONDITIONS
⊞ 48 TRANSPORT
⊞ 52 ENVIRONMENT
⊞ 56 AGRICULTURE, FORESTRY AND FISHERIES
⊞ 60 AGRI-FOODSTUFFS
⊞ 64 PRODUCTION, TECHNOLOGY AND RESEARCH
⊞ 66 ENERGY
⊞ 68 INDUSTRY
⊞ 72 GEOGRAPHY
⊞ 76 INTERNATIONAL ORGANISATIONS

Micro-thesaurus

⊟ 28 SOCIAL QUESTIONS
2806 family
2811 migration
2816 demography and population
2821 social framework
2826 social affairs
2831 culture and religion
2836 social protection
2841 health
2846 construction and town planning

Broader - Narrower - Related terms

health policy
RT health education [3211]
RT World Health Organisation [7606]
NT1 health care system
NT2 free medical care
RT social policy [2826]
NT2 private medical treatment
NT2 social medicine
NT2 standard agreement health care
NT2 State health service
RT universal health coverage [2836]
NT1 health costs

Figure 2. EuroVoc main structure

Regarding the economic domain, we have chosen the sectors Economics, Trade, Finance, Business and Competitions, Employment and working conditions, Industry as core resources.

For the ontology editing, we adopted Protégé [5], which provides a conceptual development environment and an interactive graphical tool for the design and implementation of ontologies. Because of the project implementation and validation is conducted on data provided by Italian public institutions, the Italian version of EuroVoc sectors mentioned above, was transferred into Protégé creating one class, both for each sector and for each micro-thesaurus. In order to respect the original hierarchy, every micro-thesaurus has been linked to its broader/narrower terms through the native Protégé property SubClassOf. In coherence with EuroVoc specification, the relations hasBroader, hasNarrower and isRelated have been implemented inside Protégé. The annotation process has been mainly driven by the SKOS definitions of preferredLabel, hiddenLabel, seeAlso, isDefinedBy. Some overlapping concepts were present in the three domain ontologies. In the merging phase, this was addressed by means of SKOS's mapping properties used to define links between concepts in different ontology schemes, as exactMatch, closeMatch, narrowMatch, broadMatch and relatedMatch properties.

A. *Health domain ontology*

The main objective of Health domain ontology is to provide information about chronic diseases, with focus on cardiovascular disease and address 1) general questions coming from citizens, 2) specific questions coming from health system actors. In addition to EuroVoc, other specialized resources have been accessed in order to meet specific domain requirements, in particular the Unified Medical Language System [6] meta-thesaurus, a repository of biomedical concepts and the Medical Subject Headings [7] thesaurus, often used in document indexing. Efforts have been made to migrate the EuroVoc ‘2841 health’ micro-thesaurus (i.e., health policy, health care profession, illness, medical science, nutrition, pharmaceutical industry) and a part of MESH Heading related to the Cardiovascular Diseases ‘Tree C14’ to the standard formal web ontology language OWL [8], that became a World Wide Web

ISBN: 978-1-61208-386-5

Consortium Recommendation in 2009. This migration step was crucial to performing ontology editing through Protégé and creating an initial interoperable ontology file. Then, we identified several extensions, the main ones regard drugs and disease: the extension related to drugs implied the migration of the first two levels of Anatomical Therapeutic Chemical (ATC) classification system. The extension related to diseases implied the migration of the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). The annotation process was implemented with the support of MESH thesaurus. A final extension was carried out to model measures and indicators of health and quality of care in Tuscany, developed by Health Regional Agency (ARS Toscana). Relations have been established between concepts of the different sub-hierarchies, e.g., given a diagnosis as Acute Myocardial Infarction, the ontology provide links with indicated drugs and with related diagnosis and symptoms.

B. *Economic domain ontology*

In recent years, the economic domain has shown several points of contact and overlap with the themes of health and social services. This partly because of the negative climate resulting from the world economy crisis, with strategies on spending review, compression of costs of health care, the links between income (individual and family) and access to health and social services [9]. These issues represent crucial points in the agenda of policy makers, due to the structural traits of economic crisis and it will become even more important to allocate resources in an optimal way, working on recovery aspects based on a cognitive framework in which the economic aspects will necessarily communicate with the aspects more directly related to the supply of services.

The economic domain ontology has been based on EuroVoc and some extensions have been identified. Primarily, metadata and the ontology of the Linked Open IPA public network and cooperation [10]. This resource contains the index of the national Public Administrations and the index of companies wholly owned by public authorities or with majority public capital included in the consolidated statement of public administration, as identified by the National Institute for Statistics (ISTAT). Secondly, the ISTAT Statistical Glossary [11] that provides classification and definitions of indicators and statistical terms used in the most open databases available in Italy. The final ontology has provided extension of several EuroVoc main sectors: sector 16 Economics, sector 20 Trade, sector 24 Finance, sector 40 Business and competition, sector 44 Employment and working conditions and sector 68 Industry.

C. *Social domain ontology*

Starting from the preliminary need to define the contours of the "social" in accordance with the objectives of ODINet, we have made an initial assessment of the Eurovoc resource. The nature of the instrument created to classify texts and documents could restrict the construction of an ontology that will mainly serve to classify and correctly interpret statistical data. So, we also considered the classification system defined by the United Nations Economic Commission for Europe (UNECE), used as a base to build the structure of the Statistical Data and Metadata eXchange (SDMX) guidelines for the thematic classification of official statistics. UNECE has developed a classification of activities of statistics production (Classification of statistical activities - CSA 2009) based on three levels, the first of which consists of 5 main domains 1) Demographic and social statistics, 2) Economic statistics, 3) Environment and multi-domain statistics, 4) Methodology of data collection processing, dissemination and analysis, and 5) Strategic and managerial issues of official statistics. Within the domain Demographic and social statistics, we selected the three sectors Population and Migration, Social Protection and Social Policy and other community activities; within the domain Environment and multi-domain statistics, we chose the sub-sector living conditions, poverty and cross-cutting themes. We then entered all the specific concepts used in the production of official statistics, identifying them from ISTAT and comparing with Eurostat articulation [12] [13] [14]. The definitions have been built with the aid of additional glossaries and thesauri, such as United Nation Common Database [15] and the International Statistical Institute Multilingual Glossary [16]. Subsequently, we have enriched the concepts and definitions drawn from the statistical glossaries and we have extended the theme of social protection by referring to the European system of integrated social protection statistics (ESSPROS), with its annotations and relations. Preferred and alternative labels and definitions have been incorporated by reference to EuroVoc and to terminology and correlations themes taken from the "Thesaurus for the Social Sciences" developed by the Leibniz Institute for the Social Sciences.

III. TECHNOLOGICAL PLATFORM

In this section, we describe ODINet technological platform main components.

A. *Data Harvesting*

We designed and developed a complex module able to interface with existing portals and to find accessible datasets in the web. The module, can be periodically scheduled and manages to automatically import meta and content information from a wide variety of formats, such as CSV, XLS, MDB, DBF, SHP and RDF. Data were drawn mainly from *I.Stat* (database of statistics produced by ISTAT), *dati.toscana.it* (an open data platform developed by the Tuscany Region) and *dati.gov.it* (an open data portal developed by the Italian Ministry for Public Administration and Innovation) portals. Stakeholders who joined the project and provided data for the validation scenario are *ARS Toscana* (the Regional Health Agency), *IRPET* (the Regional Institute Planning Economic of Tuscany) and *Rete Osservatori Sociali Regione Toscana* (a Social Observers Network in Tuscany).

B. *Data Indexing*

A search graph was built as main support for the search engine by combining the concepts identified during the ontology-building process and adding relationships between them according to ontologies' predicates. Since the graph was sparse, two procedures were designed and implemented to improve its semantic information. Finally, the data component was added to the graph and linked to the concepts contained in it. In such a way, the complete domain knowledge is summarized in a single graph, which can be used as the basis of the reasoning mechanism of the Search Engine to answer users' queries.

1) From the ontologies to the Search Graph

In order to produce an effective and useful search engine we built a search graph combining the previously described ontologies and the data harvested. Having a weighted graph representing our model of knowledge is a key element in our project, since it allows to answer users' search queries by browsing the graph through well-known algorithms based on distance metrics, centrality of nodes and clustering coefficient, that were partially developed and tested in a previous research project [17]. We decided to store into different graph entities the concepts corresponding to the various concepts identified inside ontologies, as well as datasets (corresponding to collections of data organized in a single table), literals (corresponding to labels associated to concepts or keywords associated to datasets) and categories (corresponding to generic themes grouping datasets from the same semantic area). Each concept has at least one associated literal, corresponding to its name, but can also be associated with multiple literals, corresponding to alternatives synonyms for that concept. For example, the concept identified by the url "http://eurovoc.europa.eu/5565" has as its preferred label the literal "cyclone" but has three further associated literals, namely "hurricane", "tornado" and "typhoon". In the current graph we have ~10K concepts, ~24K literals, 22 categories and ~9K datasets. All these entities are the nodes of our graph. Afterwards, we introduced into the graph several relationships between nodes. The relationships are established on the basis of the ones identified in the domain ontologies and SKOS predicates. Each type of relationship has a specific weight in the search engine process. In the current graph we have ~100K relationships.

2) Enriching the Search Graph with semantic information: concept-concept matching.

Analysing the search graph we have described so far, we noticed that it was extremely sparse, as we had few edges compared to the number of nodes. Since the activity of finding connections between concepts that belong to different ontologies takes a long editing time even to domain experts, we explored automated methods for discovering those links. Our main goal was to identify concepts and literals semantically connected to one another: in this way, given a search string it would be possible to identify a group of concepts semantically connected to the string searched by the user, identifying not only datasets containing the string, but also the ones connected to a concept semantically related to it. To identify entities semantically connected to one another we implemented two separate procedures named *WikiSimilarityDistance* and *WikiConceptConnection*.

a) WikiSimilarityDistance

This procedure produces an estimate of the semantic distance between two concepts through an approximation of the Google Similarity Distance: this measure calculates the semantic similarity between two words (or two sentences) on the basis of the number of pages indexed by Google in which the two words (sentences) appear together, in relation to the number of pages in which they appear singularly. The formula to calculate the (Normalized) Google similarity Distance (NGD) is the following:

$$\mathrm{NDG}(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

where *f(x)* and *f(y)* are the number of hits for words *x* and *y* respectively, *f(x,y)* is the number of pages in which both *x* and *y* occur and *M* is a parameter to take into account the order of magnitude of the number of results. Since Google APIs to access a search result are not free, we approximated this measure by indexing in Solr [18] a dump of the whole Italian Wikipedia and performing search queries against the indexed Wikipedia pages for concepts and literals that are neither too general nor too specific. Couples of concepts with an NGD value below a given threshold are said to be semantically correlated. We managed to identify, through such a procedure, ~10K semantic correlations between entities, enriching our search graph with semantic connections of different weights, according to their NGD values.

b) WikiConceptConnection

Through this procedure we managed to find further semantic associations between couple of entities exploiting the items in the "See Also" section of Wikipedia pages. The procedure associates concepts and literals in our graph to Wikipedia pages with the same name, then it parses those pages and retrieves the links in the "See Also" section, which contains a list of pages that are correlated to the current one. If a concept (or a literal) with the same name of one of the pages in the "See Also" list exists, a new edge between the two correlated entities is inserted into the graph. Through such a procedure we managed to add about 3.5K further relationships to the Search Graph.

3) Inserting datasets into the Search Graph and dataset-concept matching

For each dataset, we decided to import its title (so as to identify it), its keywords (literals to which the dataset is connected) and its description (so as to get more information about its content). Then, we connected each dataset to the concepts expressed in it. Since semantic correlations between concepts were already stored into the graph, we decided to perform a purely-syntactic matching between datasets and concepts and to index all the information about a dataset using Solr [18]. We performed a Solr query for every concept in our graph, searching its associated preferred label in the title, in the keywords and in the description of the indexed datasets. A new relationship was created between a concept and each dataset in which the concept was found to be expressed. We further assigned different weights to each of those relationships according to where and how many times the concept was found in the dataset. One of our assumptions was, for example, that if a concept was found in the dataset's title the connection with the dataset is stronger than if it was found in the description. Finally, we ran a similar procedure to match datasets with literals whose names were not the same as the one of the concepts. In total we managed to add ~20K further relationships to the Search Graph.

4) *Visual representation*

The visual representation of the Search Graph is shown in Figure 3. A zoom-in snapshot of the graph is depicted on the left side: the cloud icon identifies a concept, the table one represents a dataset, the one similar to a label represents a literal and the green one identifies a category. A zoom-out snapshot of part of the graph is shown on the right.

C. Data Search Engine

This paragraph explains how the search process works, starting from a search query entered by a user up to the final output returned by the Search Engine.

1) *Full-text Search*

In this step, we aimed to find concepts, literals and datasets that are related to the search string entered by the user. A SQLServer [19] query, performed a full-text search to find entities containing all the meaningful words (articles, prepositions and common words are deleted) entered by the user. A list of datasets containing all the search words is memorized and returned in the final results subsequently. If neither concepts nor literals are found, the datasets obtained with the query is directly returned as the final result and the procedure ends.

2) *Page Ranking and Semantic Propagation with SNA*

a) 1st propagation

In this step, we aimed to find concepts and literals semantically connected with the ones returned by the previous phase. For this purpose, we implemented a variant of the PageRank algorithm [20] to identify concepts and literals most strongly connected with the preceding ones. This algorithm executes a propagation on the search graph by taking into account the strength of a link to an entity (represented by its associated weight) in order to determine a rough estimate of how important the entity is. This step produced a list of concepts and literals semantically connected with the search query.

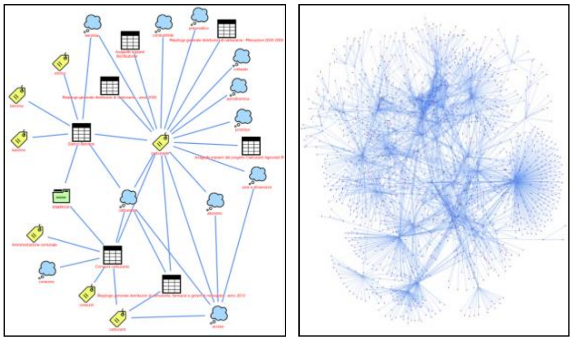


Figure 3. Visual representation of the Search Graph

b) 2nd propagation

We used the PageRank algorithm to find datasets directly connected to literals and concepts returned by the previous step. At the end we got a list of datasets connected to concepts related to the search query.

3) *Final Results*

The final result returned to the user is the union between the datasets identified with the full-text search and those returned at the end of the second propagation phase. The datasets are ordered by their rank value, a measure calculated during the search process that estimates the dataset's relevance for the search query. The flowchart of the search engine process is shown in Figure 4.

IV. RESULTS

As a first result, the ontologies may function independently and by themselves as separate knowledge basis. They provide about 4000 distinct concepts, corresponding relations and labels for synonyms.

In order to validate the overall research, we defined some contexts and use cases based on real issues identified with stakeholders. The *Caring for the elderly* use case is shown in Figure 5. The list of concepts found by the semantic search is Zonal user rate for elderly home care, Health and social care for the elderly, Taking care of elderly people for the health service, Caring for the elderly, Elderly person, Gerontology, Elderly people, Family solidarity, Pension scheme, Care allowance, Retired person, Home help, Facilities for the disabled, Older worker.

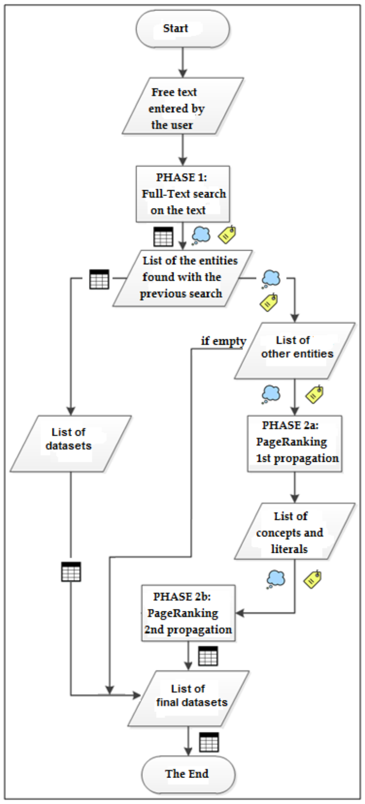

Start
Free text entered by the user
PHASE 1: Full-Text search on the text
List of the entities found with the previous search
if empty
List of other entities
List of datasets
PHASE 2a: PageRanking 1st propagation
List of concepts and literals
PHASE 2b: PageRanking 2nd propagation
List of final datasets
The End

Figure 4. Flowchart of the search engine process

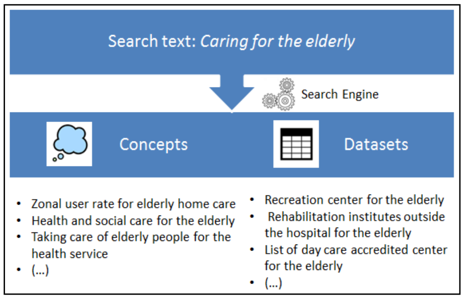

Search text: *Caring for the elderly*
Search Engine
Concepts
Datasets
• Zonal user rate for elderly home care
• Health and social care for the elderly
• Taking care of elderly people for the health service
• (...)
• Recreation center for the elderly
• Rehabilitation institutes outside the hospital for the elderly
• List of day care accredited center for the elderly
• (...)

Figure 5. Use case *Caring for the elderly*

The top five datasets returned are Recreation center for the elderly, Rehabilitation institutes outside the hospital for the elderly, List of day care accredited center for the elderly, Social services for the elderly and List of residential structures for the elderly. A dynamic web interface which displays the search results organized in semantic clusters is under testing: the user will be able to see graphical representations of the results, to select only those datasets that are more strongly connected with his search (making a dynamic disambiguation) and to navigate the associated graph.

Our platform has a number of strengths. Firstly, our search engine is both reliable, being based on well-known SNA algorithms working on graphs, and innovative, being the ontologies and the numerical datasets included in a single tool that constitutes the knowledge base of the whole system. Further, the knowledge representation's component and the search engine's module are decoupled, guaranteeing a high level of reuse and adaptivity. In fact, with proper changes to the underlying ontology, the tool can be used in completely different contexts with respect to the current ones. Moreover, the harvest procedure can be periodically scheduled and manages to automatically import data and to index the newly found datasets.

Another innovative aspect of our tool is to extend the semantic component of the system to the whole knowledge base. Our SSE is based on an ontological graph of entities, among which there are generic concepts and datasets, semantically connected to one another. This fact lets the SSE to identify datasets relating not only to the search query but also to concepts semantically related to those contained in the search text. Lastly, differently from other systems, the user can directly interact with the tool, making a dynamic disambiguation on the results and identifying those datasets that are more strongly connected to his search.

V. Discussion and Future Developments

In this paper, we have mainly presented the semantic component of the ODINet project (http://www.odinet.sister.it), which provides an innovative technological framework for data search engine.

Specifically, we have first created a unified ontology which models the conceptual hierarchies that describe the major aspects of social, economic and health fields and the main connections between them. We have designed and developed a complex module able to interface with existing portals and to find accessible datasets in the web. A search graph was built as main support for the SSE by combining the concepts identified during the ontology-building process and adding relationships between them according to ontologies' predicates. A dynamic web interface was developed to display the search results in an intuitive way, allowing the users to play a part of the answers and perform disambiguation.

The validation phase of the project is still in progress. Different stakeholders offered validation scenarios on thematic portals, correspondent to ODINet domains. The tests carried out so far have shown that our platform achieves higher performances on those portals than on generic datasets: on a number of test cases performed so far, we found that more than 80% of the datasets returned by the SSE are actually related to the search query. This percentage rises to about 90% for queries related to the ODINet's domain ontologies. In fact, our search graph was built combining thematic ontologies containing specific concepts that belong to those thematic areas.

Our system also has some weak points, among which we found that connections between concepts and datasets are not always strong enough. This mainly depends on the level of detail of the meta information associated to the datasets. Further, ODINet ontology is in Italian. However, as EuroVoc and every used resource are also available in other languages, an English version can be provided with moderate effort in the future.

Moreover, in the next few months, the technologies implemented by the ODINet data search engine will be integrated in "StatPortal Open Data" (http://www.opendata.statportal.it), an open source platform for developing open data portals.

Acknowledgment

This work is supported by the Italian Tuscany Regional Operational Programme - 2007-2013 European Regional Development Fund (ODINet project D57I12000710007).

References

[1] S. Gedzelman, M. Simonet, D. Bernhard, G. Diallo, and P. Palmer, "Building an Ontology of Cardio-Vascular Diseases for Concept-Based Information Retrieval", Computers in Cardiology, vol. 32, 2005, pp. 255−258.

[2] Office for Official Publications of the European Communities, "Thesaurus EUROVOC" Annex to the index of the Official Journal of the EC, Luxembourg Publications of the European Communities, 1995.

[3] R. Steinberger, M. Ebrahim, and M. Turchi, "JRC EuroVoc Indexer JEX - A freely available multi-label categorisation tool" Proc. LREC'2012, May 2012, pp. 798-805.

[4] N. Ivezic, A. Farhad, K. Khosrow, and B. Kulvatunyou, "Ontological Conceptualization Based on the Simple Knowledge Organization System (SKOS)", Journal of Computing and Information Science in Engineering, doi:10.1115/1.4027582, vol. 14, issue 3, May 2014, pp. 11.

[5] M. Horridge, "A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools", The University Of ManchesterEdition 1.3, March 2011

[6] O. Bodenreider, “The Unified Medical Language System (UMLS):integrating biomedical terminology”, Nucleic Acids Res. doi:10.1093/nar/gkh061, vol. 32, 2004, pp 267–270

[7] W. D. J. Stuart, J. Nelson, and B.L. Humphreys, “Relationships in Medical Subject Headings (MeSH).” National Library of Medicine,Bethesda, MD, USA, 2002. Available from: http://www.nlm.nih.gov/mesh/meshrels.html [retrieved: 12, 2014]

[8] D. L. McGuinness, and F. van Harmelen, “OWL Web Ontology Language Overview”. W3C Recommendation, February 2004. Available from http://www.w3.org/TR/2004/REC-owl-features-20040210/ [retrieved: 12, 2014]

[9] World Health Organization on behalf of the European Observatory on Health Systems and Policies, “Health, health systems and economic crisis in Europe: impact and policy implications”, 2013. Available from http://www.euro.who.int [retrieved: 12, 2014]

[10] Italian Public system of connectivity and cooperation, “Guidelines for semantic interoperability through linked open data”, 2012 Agency for Digital Italy

[11] M. Frustaci, “Glossary Economic-Statistical Multilingual”. Italian National Institute for Statistics ISTAT, Doc 8/2004, Available from http://www.istat.it [retrieved: 12, 2014]

[12] European Commission - Eurostat, “Eurostat: The Statistic Explained Glossary”. Available from http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Thematic_glossaries [retrieved: 12, 2014]

[13] European Commission - Eurostat, “RAMON Eurostat's Metadata Server”. Available from http://ec.europa.eu/eurostat/ramon/index.cfm [retrieved: 12, 2014]

[14] European Commission - Eurostat, “Coded - The Eurostat concepts and definitions database”. Available from http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL_GLOSSARY&StrNom=CODED2&StrLanguageCode=EN [retrieved: 12, 2014]

[15] United Nations Statistics Division, “United Nations Common Database – Methods and classifications”. Available from http://unstats.un.org/unsd/methods.htm [retrieved: 12, 2014]

[16] The International Statistical Institute ISI, “The multilingual ISI glossary of statistical terms”. Available from http://isi.cbs.nl/glossary/ [retrieved: 12, 2014]

[17] M. Baglioni, S. Pieroni, F. Geraci, S. Molinaro, M. Pellegrini, and E. Lastres, “A New Framework for Distilling Higher Quality information from Health Data via Social Network Analysis”. 13th International Conference on Data Mining (ICDMW.2013) IEEE, December 2013, pp. 48-55, DOI 10.1109/.142.

[18] T. Grainger and T. Potter, “Solr in Action”. Manning Publications, 2014.

[19] R. Mistry and S. Misner, “Introducing Microsoft SQL Server” 2014. Microsoft Press, 2014.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank Citation Ranking: Bringing Order to the Web”, Technical report, Stanford Digital Library Technologies Project,1998