Interoperability and Semantic Technologies 2015-16

Last Updated Tuesday, 05 July 2016

Syllabus

The course starts arguing about the difference between pair-wise system integration and multi-later system interoperability. It introduces the levels at which the interoperability problem has to be attacked using the HL7 (specifically, the Reference Information Model [1] and the Clinical Document Architcture [2]) as a case study. It discusses the two ways interoperability can be achieved: standards and translations. It shows how semantic technologies allowed translation-based interoperability on the Web.

Then, it presents semantic technologies for interoperability. These technologies are very important nowadays because the allow to treat the "variety" dimension of Big Data. It introduces RDF [3] - a flexible data model to (virtually) represent heterogenous data. It describes OWL [4] - a flexible ontological language to model heterogenous data sources in an open world where information cannot be assumed to be complete. It illustrates SPARQL [5] - a query language for RDF. It shows how to put all the pieces together in order to achieve interoperability among heterogenous information systems using principles from ontology-based data integration [6]. To close this part, R [7] is introduced as a flexible language/environment for statistical computing and the integration with SPARQL is illustrated.

The second part of the course covers the realm of interoperability among systems that process streaming data. These systems are very important nowadays because the allow to treat the "velocity" dimension of Big Data. It motivates the problem statement showing the importance for many Big Data analysis to process data stream from Sensor Networks and Social Media sources. It illustrates how to practically deal with data streams using Esper and the Event Processing Language (EPL) [8] and two solutions for real-time Big Data processing (spark [9] and flink [10]). Finally, it goes through the concepts illustrated in the first part of the course showing how semantic technologies can be extended to cope with the streaming nature of those data sources. It shows how to extend RDF to model RDF streams, how to extend SPARQL to continuously process RDF streams and how to reason on those RDF Streams [11]. Last, but not least, it illustrates with CitySensing [12] how to use all those ingredients together to fuse city data streams from multiple sensor and social sources.Exam

The exam consist in a practical part (40% of the grade) and a theoretical part (60% of the grade) [tmp dropbox link] Theoretical part

The theoretical part will be evaluated with a written and (optionally) an oral test. The written test is composed of questions to be answered in free text, regarding any of the course subjects, and excercises, regarding the more technical content. The oral test consists of a discussion about the written test and the practical part of the exam. It can include also questions on any subject of the course.Pratical part

The practical part consist in solving a realistic interoperability problem. The students will be given real heterogenous datasets (i.e., those released open by the Telecom Italia Big Data Challenge 2014 [13] and others of their choice), they will have to define a continuous information need and use the technologies illustrated in the course (see hereafter) to satisfy such a need executing queries that span the datasets. How to open your project

Before starting the project you should make sure that it is approaved.

To do so, please,
- use the following form to propose your information need and name the components of your group: http://bit.ly/IST-FORM
- Notify the submission to riccardo -dot - tommasini @ polimi -dot- it
- wait for his comments
- interact with him until you get the project approvedHow to submit your project for the exam

7 days before the exam you have to submit your project work using this form. After the submission each member of the group has to provide feedbacks using the other form. The feedback provided is not part of the evaluation, but it is required to obtain the final grade. Tools to use
- a database to store the data (students' choice)
- protege Desktop v5 to model/extend an ontology
- ontop protege plug-in 1.18 to model mappings and issues SPARQL queries to thedatabase where the data are stored
- Triplewave to create RDF streams
- RSP-services for C-SPARQL 0.5 to register C-SPARQL queries and oberve their resultsOptional tools
- Frappe-Lite ontology

- ontop 1.18 stand alone Frequently Asked Questions

Q: shall I use all the data ?
A: no, I can use subsets of the data, especially during the design and testing phase.

Q: do I have to reuse an existing ontology?
A: no, but I appreciate if you do (e.g., FraPPE-Lite or SIOC)

Q: shall I model a comprehensive ontology for the data I choose?
A: no, you can model in the ontology just the terms you need to satisfy the information need you choose.

Q: shall I build mappings for every single data item in the data source or term in the ontology?
A: no you can build just the mappings you need to satisfy the information need you choose.

Q: some data sources are accessible using APIs, shall I integrate the APIs?
A: no, you can just download some data and treat them as a file

Q: some data are in JSON, what shall I do?
A: you may want to cover them in CSV or TSV e.g., using a command line tool as json2csv

Q: I have latitude and longitude of a geo-point, how can I obtain the ID of the cell of the grid?
A: you can use the following codeminLat=45.356686
minLon=9.011491
cellHeight=0.00211101
cellWidth=0.00301197
verticalIndex = (int)((lat - minLat) / cellHeight);
horizontalIndex = (int)((lon - minLon) / cellWidth);
cell_ID = (int)(((verticalIndex * 100) + horizontalIndex) + 1);LecturesHereafter, you find the tentative calendar of the course. The material presented in class will be linked here and posted on twitter on @manudellavalle.

- 7.3.2016 - V.S8-B - Introduction [slideshare]
- 14.3.2016 - V.S8-B - HL7 from syntax (v2.x) to semantics (v3.x) [slideshare], the Reference Information Model (RIM) [pdf (to study), slideshare (original exteded version)] and the Clinical Document Archtiecture (CDA) [pdf (to study), slideshare (original extended version)] as cases of sematic interoperability

- 21.3.2016 - V.S8-B - Semantic Web technologies [slideshare]
- 23.3.2016 - V.S8-B - RDF [slideshare] and solution of the exercise proposed in class [whiteboard,txt]

- 30.3.2016 - V.S8-B - OWL [tmp dropbox link] and Protégé [link] practice session [tmp dropbox link]
- 4.4.2016 - V.S8-B - SPARQL basics [tmp dropbox link]
- 11.4.2016 - V.S8-B - SPARQL in class exercise solutions [tmp dropbox link]

- 20.4.2016 - V.S8-B - Semantic annotation of HTML pages and Schema.org [tmp dropbox link]

- 27.4.2016 - V.S8-B - R2RML [tmp dropbox link]
- 2.5.2016 - V.S8-B - Putting it all together (please download and install protege 5.0 beta and H2) [tmp dropbox link]

- 4.5.2016 - V.S8-B - finishing the lesson on putting it all together and Q/A
- 9.5.2016 - V.S8-B - It's a streaming world [tmp dropbox link] Stream and Complex Event Processing [tmp dropbox link]

- 11.5.2016 - V.S8-B - Event Processing Language [tmp dropbox link]

- 16.5.2016 - V.S8-B - Real-time Big Data [map-reduce explained visually] with Spark and its applications [slideshare (slides 1-35 only), pdf (slides 12-22), example application using Twitter, usb stick with the code (2.2GB), tutorial material]

- 23.5.2016 - V.S8-B - Real-time Big Data with Flink and its applications [tmp dropbox link, code examples]

- 25.5.2016 - V.S8-B - Taming Velocity and variety simultaneously with Stream Reasoning [tmp dropbox link]

- 30.5.2016 - V.S8-B - City Sensing use case [BIS keynote on slideshare]

- 6.6.2016 - V.S8-B - RDF Stream Processing practice: C-SPARQL [tmp dropbox link], Triplewave and RSP services [tmp dropbox link], setup guide [tmp dropbox link]

- 8.6.2016 - V.S8-B - Exam preview and project work presentation [tmp dropbox link] overview on Semantic Web techs
[tmp dropbox link]
- 13-15-20-22.6.2016 - V.S8-B - 8 hours of supported project work

Please check here the dates before coming to the lecture room. Any change will be communicated using the maling/SMS system of PoliMI. Please, make sure your email/phone is present.References

[1] http://www.hl7.org/implement/standards/product_brief.cfm?product_id=77
[2] http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7
[3] http://www.w3.org/RDF/
[4] http://www.w3.org/TR/owl-overview/
[5] http://www.w3.org/TR/sparql11-overview/
[6] https://en.wikipedia.org/wiki/Ontology-based_data_integration
[7] https://www.r-project.org/
[8] http://www.espertech.com/esper/release-5.2.0/esper-reference/html/index.html
[9] http://spark.apache.org/streaming/
[10] https://flink.apache.org/
[11] http://streamreasoning.org/
[12] http://jol.telecomitalia.com/jolskil/tag/city-sensing/
[13] https://dandelion.eu/datamine/open-big-data/