Brainstorming about WG1
Representation of structured DS

Mihai Lupu, Raquel Trillo-Lado
TU Wien - Austria, University of Zaragoza - Spain
Outline

- Introduction
  - Topics related to WG1 according to the MoU
  - Statistical data about people on the Working Group 1 and feedback received

- Topics on which Keystone members of WG1 are currently working on

- Brainstorming: more ideas, topics, relevant people or institution to follow...
  - Related works
  - Challenges, ideas
  - Relevant people or institutions out of Keystone to follow
Introduction. Topics

WG1
Representation of Data Sources

WG2
Keyword-based Search

WG 3: User Interaction & keyword query interpretation

WG4
Research integration, showcases, benchmarks and evaluations

Data

Feedback

Happy Users or SW agents

SESSION 2. 6TH KEYSTONE MEETING
WG1: Representation of Data Sources

WG1.A: Generation of Structured Data

WG1.B: Storing & Indexing Structured Data

WG1.C: Characterization, Integration & Federation of Data Sources

WG1.D: Selection & Retrieval of Data Sources
Introduction. Statistics

- 162 members: 41 Female and 121 Males (http://www.keystone-cost.eu/keystone/work-group/wg1/)
- Most of the members are active currently: 159 members (119 M, 40 F).

<table>
<thead>
<tr>
<th>Country</th>
<th>Nº Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albania (AL)</td>
<td>1</td>
</tr>
<tr>
<td>Austria (AT)</td>
<td>2</td>
</tr>
<tr>
<td>Belgium (BE)</td>
<td>3</td>
</tr>
<tr>
<td>Bulgaria (BG)</td>
<td>3</td>
</tr>
<tr>
<td>Switzerland (CH)</td>
<td>2</td>
</tr>
<tr>
<td>Cyprus (CY)</td>
<td>1</td>
</tr>
<tr>
<td>Germany (DE)</td>
<td>12</td>
</tr>
<tr>
<td>Estonia (EE)</td>
<td>1</td>
</tr>
<tr>
<td>Greece (EL)</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country</th>
<th>Nº Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain (SP)</td>
<td>27</td>
</tr>
<tr>
<td>Finland (FI)</td>
<td>2</td>
</tr>
<tr>
<td>France (FR)</td>
<td>10</td>
</tr>
<tr>
<td>Croatia (HR)</td>
<td>4</td>
</tr>
<tr>
<td>Ireland (IE)</td>
<td>5</td>
</tr>
<tr>
<td>Italy (IT)</td>
<td>13</td>
</tr>
<tr>
<td>Macedonia (MK)</td>
<td>1</td>
</tr>
<tr>
<td>Malta (MT)</td>
<td>4</td>
</tr>
<tr>
<td>Netherlands (NL)</td>
<td>5</td>
</tr>
<tr>
<td>Norway (NO)</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country</th>
<th>Nº Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poland (PL)</td>
<td>2</td>
</tr>
<tr>
<td>Portugal (PT)</td>
<td>5</td>
</tr>
<tr>
<td>Romania (RO)</td>
<td>23</td>
</tr>
<tr>
<td>Serbia (RS)</td>
<td>7</td>
</tr>
<tr>
<td>Sweden (SE)</td>
<td>2</td>
</tr>
<tr>
<td>Slovenia (SI)</td>
<td>2</td>
</tr>
<tr>
<td>Slovakia (SK)</td>
<td>3</td>
</tr>
<tr>
<td>New Zealand (TK)</td>
<td>2</td>
</tr>
<tr>
<td>Ukraine (UA)</td>
<td>2</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>5</td>
</tr>
</tbody>
</table>
Feedback received from members of WG 1 per country:

<table>
<thead>
<tr>
<th>Country</th>
<th>Nº Members</th>
<th>Country</th>
<th>Nº Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albania (AL)</td>
<td></td>
<td>Spain (ES)</td>
<td>20</td>
</tr>
<tr>
<td>Austria (AT)</td>
<td>2</td>
<td>Finland (FI)</td>
<td>1</td>
</tr>
<tr>
<td>Belgium (BE)</td>
<td></td>
<td>France (FR)</td>
<td>2</td>
</tr>
<tr>
<td>Bulgaria (BG)</td>
<td>1</td>
<td>Croatia (HR)</td>
<td></td>
</tr>
<tr>
<td>Switzerland (CH)</td>
<td>1</td>
<td>Ireland (IE)</td>
<td>1</td>
</tr>
<tr>
<td>Cyprus (CY)</td>
<td></td>
<td>Italy (IT)</td>
<td>5</td>
</tr>
<tr>
<td>Germany (DE)</td>
<td>2</td>
<td>Macedonia (MK)</td>
<td></td>
</tr>
<tr>
<td>Estonia (EE)</td>
<td>1</td>
<td>Malta (MT)</td>
<td></td>
</tr>
<tr>
<td>Greece (EL)</td>
<td>5</td>
<td>Netherland (NL)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Norway (NO)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poland (PL)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Portugal (PT)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Romania (RO)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Serbia (RS)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sweden (SE)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Slovenia (SI)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Slovakia (SK)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>New Zealand (TK)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ukraine (UA)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>United Kingdom (UK)</td>
<td></td>
</tr>
</tbody>
</table>
Research groups involved in the feedback received:

<table>
<thead>
<tr>
<th>Country</th>
<th>Nº Members</th>
<th>Research groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Vienna B: 1</td>
<td></td>
</tr>
<tr>
<td>Switzerland (CH)</td>
<td>Geneva: 1</td>
<td>University of Geneva, Faculty of economics and social sciences, department Hautes études commerciales</td>
</tr>
<tr>
<td>Germany (DE)</td>
<td>Hannover: 2</td>
<td>L3S Research Center of the Leibniz University Hannover</td>
</tr>
<tr>
<td>Finland (FI)</td>
<td>Jyvaskyla: 1</td>
<td>Industrial Intelligent Group. MIT Department, Agora Centre. University of Jyvaskyla.</td>
</tr>
<tr>
<td>Ireland (IE)</td>
<td>Galway: 1</td>
<td>Insight Centre for Data Analytics. National University of Ireland (NUI Galway,)</td>
</tr>
</tbody>
</table>
Introduction. Statistics

- Research groups involved in the feedback received:

<table>
<thead>
<tr>
<th>Country</th>
<th>Nº Members</th>
<th>Research groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greece (EL)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thessaly: 1</td>
<td></td>
<td>Department of Computer Science and Biomedical Informatics, School of Sciences,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>University of Thessaly.</td>
</tr>
<tr>
<td>Crete A: 1</td>
<td></td>
<td>Computer Science Department, University of Crete.</td>
</tr>
<tr>
<td>Crete B: 1</td>
<td></td>
<td>Software Technology and Network Applications Laboratory, Department of Electronic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&amp; Computer Engineering, Technical University of Crete.</td>
</tr>
<tr>
<td>Athens: 1</td>
<td></td>
<td>Institute for the Management of Information Systems of the Research and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Innovation Centre ATHENA, located in Athens.</td>
</tr>
<tr>
<td>Peloponnese: 1</td>
<td></td>
<td>Knowledge and Uncertainty Research Laboratory (RAB Lab), Department of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Informatics and Telecommunications of the University of Peloponnese.</td>
</tr>
</tbody>
</table>
Research groups involved in the feedback received:

<table>
<thead>
<tr>
<th>Country</th>
<th>Nº Members</th>
<th>Research groups</th>
</tr>
</thead>
</table>
| Spain (SP)| • Segovia: 1  
• Coruña: 5  
• Santiago: 1  
• Zaragoza A: 7  
• Zaragoza B: 2  
• Málaga: 3  
• Vigo: 1 | • DataWeb Group, Department of Computer Science, University of Valladolid, Segovia.  
• Databases Laboratory (LBD), Computer Science and Technology Faculty, University of A Coruña.  
• Computer Graphics and Data Engineering (COGRADE) Singular Information Technologies Research Center (CiTIUS), University of Santiago de Compostela.  
• Computer Science and Software Engineering Department (DIIS), University of Zaragoza.  
• Aragon Institute of Engineering Research (I3A), University of Zaragoza.  
• Khaos Research, Department of Computer Languages and Computing Sciences, University of Malaga.  
• ETSE Telecomunicación, University of Vigo |
| France (FR)| • Lyon: 1  
• Paris: 1 | • University Claude Bernard Lyon  
• CNRS - Centre national de la recherche scientifique |
### Research groups involved in the feedback received:

<table>
<thead>
<tr>
<th>Country</th>
<th>Nº Members</th>
<th>Research groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy (IT)</td>
<td>• Modena: 2&lt;br&gt;• Trento A: 1&lt;br&gt;• Trento B: 1&lt;br&gt;• Bologna: 1</td>
<td>• Databases (DBGroup), University of Modena and Reggio Emilia, Modena.&lt;br&gt;• Data Management Group, Department of Information Engineering and Computer Science at the University of Trento.&lt;br&gt;• Process and Data Intelligence (PDI), Information Technology Center, at Fondazione Bruno Kessler.&lt;br&gt;• Department of Computer Science and Engineering, University of Bologna.</td>
</tr>
<tr>
<td>Romania (RU)</td>
<td>• Bucharest: 1</td>
<td>• Faculty of Automatic Control and Computers Computer Science Department, University Politehnica of Bucharest.</td>
</tr>
<tr>
<td>Estonia (EE)</td>
<td>• Tallinn: 1</td>
<td>• School of Information Technologies, Department of Software Science, Tallinn University of Tecnologies.</td>
</tr>
<tr>
<td>Bulgaria (BG)</td>
<td>• Sofia: 1</td>
<td>• Bulgarian Academy of Sciences, Sofia.</td>
</tr>
</tbody>
</table>
Introduction. Statistics

Jyvaskyla
Santiago de Compostela
Vienna
Trento A
Segovia
Coruña
Trento B
Modena
Bologna
Geneva
Zaragoza A
Sofia
Athens Thessaly
Crete B
Crete A
Hannover A
Galway
Trento A
Zaragoza B
Vienna A
Coruña

Belgrade, 21 Feb 2017
Session 2. 6th Keystone Meeting
Topics WG1 members are working on

**WG1.A**
Generation of Structured Data

**WG1.B**
Storing & Indexing Structured Data

**WG1.C**
Characterization, Integration & Federation of Data Sources

**WG1.D**
Selection & Retrieval of Data Sources

**Groups: persons**
- Santiago: 1
- Zaragoza: 3
- Vienna A: 1
- Sofia: 1
- Lyon: 1
- Crete B: 1
- Athens: 1
- Malaga: 1

**WG1.A**
- Vienna A: 1
- Vienna B: 1
- Segovia: 1
- Coruña: 5
- Jyvaskyla: 1
- EL-Greece
- Peloponnese: 1
- Bucharest: 1
- Thessaly: 1

**WG1.B**
- Galway: 1
- Santiago: 1
- Hannover: 4
- Zaragoza A: 7
- Zaragoza B: 2
- Poland: 1
- Geneva: 1
- Vienna B: 1
- Modena: 2
- Trento A: 1
- Trento B: 1
- Paris: 1
- Sofia: 1
- Bologna: 1

**WG1.C**
- Hannover: 4
- Trento A: 1
- Trento B: 1
- Paris: 1
- Bologna: 1

**Others**
All groups

BELGRADE, 21 FEB 2017
Brainstorming. Topics WG1.A

- Where structured data come from?
- Who or what generates structured data?

- From **Sensors** and **IoT devices**
- From **unstructured or semi-structured data sources**:
  - Documents written in **Natural Languages**
  - **Traditional HTML Web pages**
- From **human users** in a **collaborative way**:
  - Definition ontologies or common vocabularies or metadata
  - Population of those ontologies with instances
- From other **Structured Data sources**:
  - Translation or transformation of data
  - **Humans interactions**
  - **SW interactions**
Brainstorming. Topics WG1.A

- What data from sensors should be considered?
- How to exploit structured data considering privacy and security issues?
- Which device should process the data?
  - Fog Computing VS Cloud Computing
- How to anonymize personal data to protect people?
  - Guidelines, techniques, tools?
- What other dimensions should be considered? How to deal with the Right to Forgetfulness?
  - With provenance information?
  - Blockchains

- Raw Data VS Smart Data
  - Generating data each certain time (period) VS generating data only when a changes happens (changes bigger than a certain threshold)
  - Generating data all the time VS consulting current values of sensors when they are required (push vs pull approaches)

- Dimensions of Security and/or Safety:
  - Availability
  - Confidentiality
  - Integrity
  - Authenticity
  - Traceability (provenance)
Brainstorming. Topics WG1.A

- Non-textual Data:
  - Images
  - Video
  - Multidimensional arrays (environmental data)

- Generation of Structured Data

- How to deal with non-textual data?
- What features to consider for: images, video, streaming data from environmental sensors?
- Should all data have a unique identifier? How to build it?
- Should this kind of data be associated to a geographical position? Which granularity should be considered: country, GPS coordinates?
Brainstorming. Topics WG1.A

- Generation of Structured Data
  - Methodologies, standards, good practices to publish or generate structured data?
  - Which are the main problems of these ones?

- Methodologies to publish structured data (and/or data sources) on different medias:
  - On the Web:
    - Principles of Linked Data
    - RDFa Standard
  - Textual ETL and Web ETL
  - Entity Extraction
  - Entity De-Duplication
    - George Bush, G. Bush
    - The King of Spain, Felipe VI

- How to consume Linked Data Web? Crawling of Web pages with Web Mining considering RDFa annotations VS. Querying SPARQL endpoints?
  - Disadvantages of crawling and Web Mining/Scrapping:
    - Hidden Web: Pages are dynamically generated
  - Disadvantages of SPARQL endpoints:
    - How to express the information need in SPARQL?
Brainstorming. Topics WG1.A

- How to create ontologies or Knowledge Bases?
- How to populate them?
- Advantages and disadvantages of Bottom-Up approaches from Folksonomies VS Top-Down approaches from the knowledge of the domain experts?
- What kind of recommender systems should be used?
  - Collaborative, Content-Based, Knowledge-Based, Context Aware, Hybrid?
  - Which should recommend?

Creation of Knowledge Bases (KB): A KB consist of a set of ontologies and a set of instances consistent with the constraints defined in the ontologies:

- How to define new vocabularies or ontologies:
  - RDFS
  - OWL
- How to populate those KB:
  - RDF, OWL
  - ETL-based Systems
  - Recommendation systems to suggest attributes/properties and values to be included by users. For example Wikinfoboxer ([http://sid.cps.unizar.es/Infoboxer](http://sid.cps.unizar.es/Infoboxer) and [http://sid01.cps.unizar.es/#!/login](http://sid01.cps.unizar.es/#!/login))
Brainstorming. Topics WG1.A

- Methodologies, tools and techniques to translate Structured Data sources?

- Standards to translate Relational Databases to RDF Stores:
  - Direct Mapping (automatic way)
  - Relational 2 RDF Matching Language (R2RML) (refine way)

- Extraction, Transformation and Load (ETL)
  - Data curation and integration
    - Multilingual
    - Heterogeneity
  - Provenance information

- Materialization or not of the RDF data? i.e., Using wrappers to access the original data source by using a different models (RDF) or dealing with the data redundancy?

- How to deal with different models with different semantics: Relational model VS. RDF model?
Brainstorming. Topics WG1.A

- More about this point? Related Works, etc?
- To be filled by all WG members and collaborators or in the session

...
Introduction. Topics WG1.B

- What kind of models exists?
- What structures and indexes are used for those models?

**Types of Structured Data Sources**
- Relational Databases
- Documental Databases
- Graph-oriented databases (Triple Stores)
- Multivalued Databases
- Object-oriented databases
- Columnar Databases
- Key-value databases
- Multi-model databases

These data sources can be **distributed** on a network. Moreover, **federation** of independent data sources could be required.

- Which models are more appropriate to satisfy the following properties?
  - Atomicity, Consistency, Isolation Durability (ACID)
  - Basically Availability, Soft-State, Eventually Consistent (BASE)
  - Consistency, Availability, Partition Tolerance (CAP)

- Which are the structures and indexes to store these data sources and retrieve data from them? Which languages apart from SQL and SPARQL?
### Introduction. Topics WG1.B

**Storing and Indexing Structured Data**

- What kind of models exist?
- What structures and indexes are used for those models?

- Structures and Indexes to store Structured Data Sources:
  - In Memory VS In Disk

  **Compact** structures VS structures over plain data

- **Self-Indexing Structures.** Both the index and the data are kept in a unique in-memory data structure that allows indexed searches and to recover the original data.

- Any examples?
  - Relational Databases:
    - Balance trees and B+ trees
  - Documental Databases:
    - Inverted index
  - Several members of Keystone have proposed different structures for RDF:
    - Head Dictionary Triple (HDT), HDT-MapReduce, HDT++
    - RDFCSA (A compact RDF store based on compressed Suffix Arrays, a well known self-index)
    - Compressed vertical partitioning for RDF (K2-triples)
Introduction. Topics WG1.B

- How to deal with the dynamic nature of RDF or other kind of Structured Data Sources?
- How to deal with the different versions and temporal data?

- Structures and indexes to deal with the evolution along time of Structured Data Sources

- Any examples?
  - Several members of Keystone has proposed different structures for versioning RDF:
    - RDF-Archive or v-RDFCSA (based on RDFCSA)
    - Compressed Suffix-Array for Temporal-Graph
    - Compressed kd-tree for temporal-Graph
  - Other Keystone members have defined ontologies and query languages to deal with versions of RDF
    - Ontology for temporal reasoning based on Extended Allen’s Interval Algebra
    - Query Language for Multi-version Data Web Archives
Introduction. Topics WG1.B

- More about this point? Related Works, etc?
- To be filled by all WG members and collaborators or in the session

- ...

Storing and Indexing Structured Data

BELGRADE, 21 FEB 2017

SESSION 2. 6TH KEYSTONE MEETING
Introduction. Topics WG1.C

- Which metadata should be considered to describe a (RDF) data source?
- How to evaluate the quality of a Data Source?

- Metadata describing Data Sources
  - Standards to describe Data Sources:
    - RDFS, OWL
    - DCAT (https://www.w3.org/TR/vocab-dcat/)

- Measurements of the Quality of Data Sources
  - Metrics considered:
    - Graph Connectivity

- Methods and tools
  - qSKOS, Skosify, PoolParty

- Some Keystone members are working on a survey paper about “DataSet Profiling”

- Some Keystone members are working on:
  - “Quantifying the connectivity of a semantic warehouse”
  - “Automatic methods to report the quality of thesauri or other sources represented in SKOS”
  - “BEAUFORD: A Benchmark for Evaluation of Formalization of OWL Definitions”
Introduction. Topics WG1.C

- More about this point? Related Works, etc?
- To be filled by all WG members and collaborators or in the session

Characterization of Data Sources

- ...

BELGRADE, 21 FEB 2017

SESSION 2. 6TH KEYSTONE MEETING
Introduction. Topics WG1.D

Selection & Retrieval of Data Sources

- How to discover structure (RDF) data sources?
- How to integrate/fusion structure (RDF) data sources?
- Recommendation of new data sources by considering the dataset profiles of the data sources already used or consulted by a user.
- Ontology Matching techniques to automatically discover equivalent concepts, properties and instances from two different ontologies (or KB).

- Some Keystone members are working on a survey paper about “DataSet Profiling and Recommendation” and “Intension-based DataSet Recommendation for Data-Linking”.
- Some Keystone members are working on “Generating Benchmark Data for Entity Matching” (EMBench).
Introduction. Topics WG1.D

- More about this point? Related Works, etc?
- To be filled by all WG members and collaborators or in the session

Selection & Retrieval of Data Sources

- ...

SESSION 2. 6TH KEYSTONE MEETING
List of Participants in WG 1 who provide feedback

- Dr Stefan Dietze
- Prof Gilles Falquet
- Prof Antonio Fariña Martínez
- Dr Francesco Guerra
- Dr Claudia Ifrim
- Dr Mihai Lupu
- Prof José Ramón Ríos Viqueira
- Dr Tarcísio Souza
- Prof Vagan Terziyan
- Dr Raquel Trillo-Lado
- Prof Yannis Velegrakis
- Dr Manolis Wallace
- Dr Sergio Ilarri
- Dr Velislava Stoykova
- Dr Enn Õunapuu
- Dr Jonh Breslin
- Dr Laura Po
- Dr Elena Demidova
- Dr Genoveva Vargas
- Dr Marín López Nores
- Dr Javier Nogueras
- Dr Guilles Falquet
- Dr Ana Cerdeira-Pena
- Dr Ramón Hermoso
- Ángel Luís Garrido
- Dr Mª del Mar Roldan
- Dr Fernando Bobillo
- Dr Carlos Bobed
- Dr Eduardo Mena
- Dr Javier Lacasta
- Catarina Ferreira da Silva
- Dr.Pablo Fafalios
- Dr Ilaria Bartolini
- Dr Ekaterini Ioannou
- Dr Yannis Stavrakas
- Dr Mauro Dragoni
- Dr Vagan Terzivan
- Dr Ioannis Anagnostopoulos
- Dr José Ramón Paramá Gabía
- Dr Javier D. Fernández
- Dr Miguel A. Martínez
- Prof José F. Aldana
- Prof Nieves R. Brisaboa
- Dr Susana Ladra