Towards a semantic-based representation and computation of quantitative indexes for quality management of requirements. The RDFIndex approach.

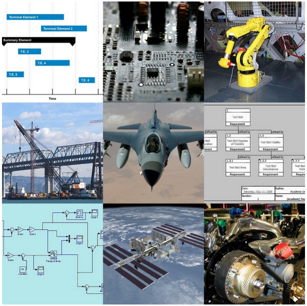
Jose María Álvarez-Rodríguez (speaker)
and
{Juan Llorens, Manuela Alejandres and
Jose Miguel Fuentes}

INCOSE 2014 | 24th Annual INCOSE International Symposium
Las Vegas, USA

The Motivating example...

Let's suppose that...

1. We want to create a "Requirements Quality index"...
2. ...to know which is the **CCC** (Correctness, Consistency and Completeness) of our specification.
3. ...to collect in just one value a set of indicators.
4. ...to make changes in our product or system measuring the impact.



If it fails, people die.
Quality is a MUST.

The Motivating example...

We already have some guidelines and tools...

- We have the "**INCOSE's Guide for Writing Requirements**" .
- ...or the ISO Quality Model.
- ...or the quality metrics in the **Requirements Quality Analyzer** of The Reuse Company (http://www.reusecompany.com/requirements-quality-suite/66-requirements-quality-analyzer-rqa).
- ...

The Motivating example...

Benefits of using an index...

1. Creation of valuable data and information.
2. Generation of know-how to make some policy.
3. Re-use of a great effort and commitment by experts in some area.
4. Rank entities according to a quantitative value.
5. ...

The Motivating example...

Drawbacks of existing indexes...

1. **Data heterogeneity**: different datasources, formats and access protocols.
2. **Structure**: math models to aggregate some indicators that can change over time.
3. **Computation process**: observations are gathered and processed, *somehow*, generating a final value.
4. **Documentation**: multilingual and multicultural character of information.
5. ...

..that imply the necessity of...

1. **Accessing data** and information under a **common and shared data model**.
2. **Representing** the evolving **structure of the index**.
3. **Computing** the index to improve transparency.
4. Providing **context-aware documentation**: user-profile.
5. ... **Exploiting valuable data and metadata**.

The Motivating example...

Drawbacks of existing indexes...

1. **Data heterogeneity**: different datasources, formats and access protocols.
2. **Structure**: math models to aggregate some indicators that can change over time.
3. **Computation process**: observations are gathered and processed, *somehow*, generating a final value.
4. **Documentation**: multilingual and multicultural character of information.
5. . . .

..that imply the **necessity** of...

1. **Accessing data** and information under a **common and shared data model**.
2. **Representing** the evolving **structure of the index**.
3. **Computing** the index to improve transparency.
4. Providing **context-aware documentation**: user-profile.
5. ... **Exploiting valuable data and metadata**.

...but...Is it a common problem?

...some indexes (per domain)...

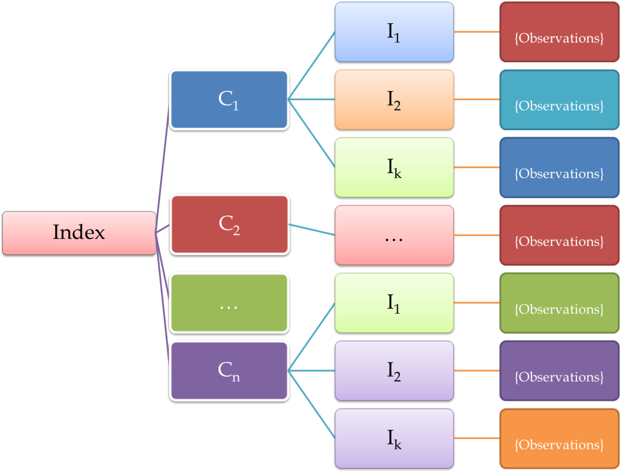
1. Bibliography: the JCR and JSR, etc.
2. Government: the GDP, etc.
3. Web: the Webindex, etc.
4. Health: the "Health Index", the "Ocean Health Index", etc.
5. Cloud: the CSC Cloud Usage Index, the VMWare index, the SMI index, etc.
6. ...to name a few per domain and creators.

A quantitative index is...

1. It is technique to **collect in just one value** a set of indicators.
2. It can be divided into: index, component and indicator.
3. An index is calculated by aggregating n components using an **OWA operator** [1].
4. A component is also calculated by aggregating n indicators using an OWA operator.
5. Observations are aligned to an indicator including some metadata such as: location, measure, value, etc.

[1] **Ordered weighted averaging** operator: $\sum_{i=1}^{n} w_i a_i$, where w_i is the weight of the aggregated element a_i.

Graphical view of a quantitative index...



Statistics and the Web of Data

Vocabularies

1. The Statistical Core Vocabulary [5] (SCOVO), a former standard to describe statistical information in the Web of Data (2009).
2. The RDF Data Cube Vocabulary [2], an adaptation of the ISO standard SDMX (Statistical Data and Metadata Exchange Vocabulary) (2013).
3. The DDI-RDF discovery vocabulary, a metadata vocabulary for documenting research and survey data (2013) [1] .

Preliminary evaluation...

Existing RDF-based vocabularies enable us the possibility of modelling and representing statistical data.

Statistics and the Web of Data

Vocabularies

1. The Statistical Core Vocabulary [5] (SCOVO), a former standard to describe statistical information in the Web of Data (2009).
2. The RDF Data Cube Vocabulary [2], an adaptation of the ISO standard SDMX (Statistical Data and Metadata Exchange Vocabulary) (2013).
3. The DDI-RDF discovery vocabulary, a metadata vocabulary for documenting research and survey data (2013) [1] .

Preliminary evaluation...

Existing RDF-based vocabularies enable us the possibility of modelling and representing statistical data.

Statistics and the Web of Data

Statistics and Linked Data

1. "Defining and Executing Assessment Tests on Linked Data for Statistical Analysis" (2011) [8].
2. "Publishing Statistical Data on the Web" (2012) [7].
3. "Publishing open statistical data: the Spanish census" (2011) [4].
4. "Publishing Statistical Data following the Linked Open Data Principles: The Web Index Project" (2012) [6].
5. "Linked Open Data Statistics: Collection and Exploitation" (2013) [3].
6. Some works in the "RDF Validation Workshop 2013" (`http://www.w3.org/2012/12/rdf-val/`).

Preliminary evaluation...

All of the approaches are/were focused on data publishing/consumption...but...

- **Validation** of statistical data and/or structure...and
- the **Computation** process are still **open issues**.

Statistics and the Web of Data

Statistics and Linked Data

1. "Defining and Executing Assessment Tests on Linked Data for Statistical Analysis" (2011) [8].
2. "Publishing Statistical Data on the Web" (2012) [7].
3. "Publishing open statistical data: the Spanish census" (2011) [4].
4. "Publishing Statistical Data following the Linked Open Data Principles: The Web Index Project" (2012) [6].
5. "Linked Open Data Statistics: Collection and Exploitation" (2013) [3].
6. Some works in the "RDF Validation Workshop 2013" (`http://www.w3.org/2012/12/rdf-val/`).

Preliminary evaluation...

All of the approaches are/were focused on data publishing/consumption...but...

- **Validation** of statistical data and/or structure...and
- the **Computation** process are still **open issues**.

Main Contributions

1-Representation

A **high-level model on top of the RDF Data Cube Vocabulary** for representing quantitative indexes.

2-Computation

A **Java-SPARQL based processor** to exploit the meta-information, validate and compute the new index values.

Main Contributions

1-Representation

A **high-level model on top of the RDF Data Cube Vocabulary** for representing quantitative indexes.

2-Computation

A **Java-SPARQL based processor** to exploit the meta-information, validate and compute the new index values.

Example: Building the "The Naive Requirement Quality Index".

- Components: **"Maintainability"** (c_1) and **"Usability"** (c_2).
- Indicators: **"Stability"** (in_1) and **"Understandability"** (in_2).
- The index, i, is calculated through the **ordered weighted averaging** (OWA) operator: $\sum_{i=1}^{n} w_i c_i$, where w_i is the weight of the component c_i
- All **observations** must be **normalized** using the *z-score* before computing intermediate and final values for each indicator, component and index.

Example of indicator observations from the WorldBank.

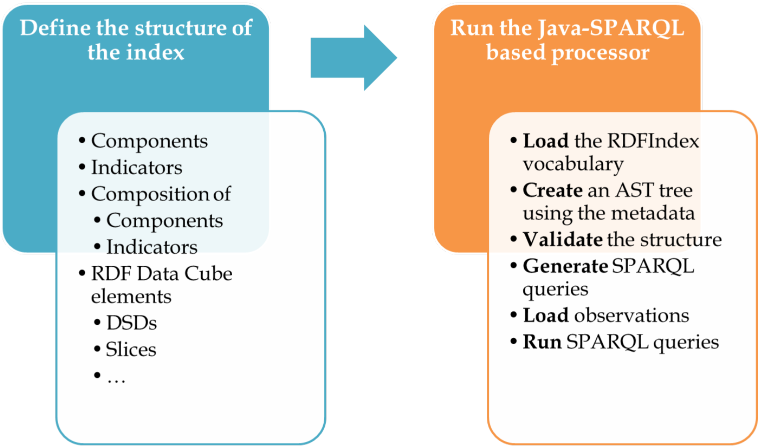
Description	Requirement	Subsystem	Value	Status
Stability	R_1	S_1	0.5	Normal
Understandability	R_1	S_1	0.8	Normal
Stability	R_2	S_1	0.3	Normal
Understandability	R_2	S_1	0.7	Normal
...	...	...	...	...
I_i	R_j	S_k	v	$status_z$

Representing and computing data with the RDFIndex (summary)

Steps

1. Define the structure and computation of the index with the RDFIndex vocabulary.
2. Run the Java SPARQL based processor:
    - Load the RDFIndex ontology to have access to common definitions.
    - Create a kind of Abstract Syntax Tree (AST) containing the defined meta-data (our index).
    - Validate the structure with an AST walker (structure and RDF Data Cube normalization).
    - Create the SPARQL queries to compute the index (other AST walker).
    - Load the observations and run the SPARQL queries to generate new observations.

Graphical view of the RDFIndex workflow...



Underlying definitions

Observation-o

It is a tuple $\{v, m, s\}$, where v is a numerical value for the measure m with an status s that belongs to only one dataset of observations O.

Dataset-q

It is a tuple $\{O, m, D, A, T\}$ where O is a set of observations for only one measure m that is described under a set of dimensions D and a set of annotations A. Additionally, some attributes can be defined in the set T for structure enrichment.

Aggregated dataset-aq

It is an aggregation of n datasets q_i (identified by the set Q) which set of observations O is derivate by applying an OWA operator p to the observations O_{q_i}.

Scope and Consequences

Necessary condition for the computation process

An aggregated dataset aq defined by means of the set of dimensions D_{aq} can be computed iif $\forall q_j \in Q : D_{aq} \subseteq D_{q_j}$. Furthermore the OWA operator p can only aggregate values belonging to the same measure m.

- The set of dimensions D, annotations A and attributes T for a given dataset Q is always the same with the aim of describing all observations under the same context.
- An index i and a component c are aggregated datasets. Nevertheless this restriction is relaxed if observations can be directly mapped to these elements without any computation process.
- An indicator in can be both dataset or aggregated dataset.
- All elements in definitions must be uniquely identified.
- An aggregated dataset is also a dataset.

Mapping the RDFIndex to the RDF Data Cube Vocabulary

Concept	Vocabulary element	Comments
Observation o	`qb:Observation`	Enrichment through annotations
Numerical value v	`xsd:double`	Restriction to numerical values
Measure m	`qb:MeasureProperty` `sdmx-measure:obsValue`	Restriction to one measure
Status s	`sdmx-concept:obsStatus`	Defined by the SDMX-RDF vocabulary
Dataset q	`qb:dataSet` and `qb:qb:DataStructureDefinition`	Metadata of the `qb:dataSet`
Dimension $d_i \in D$	`qb:DimensionProperty`	Context of observations
Annotation $a_i \in A$	`owl:AnnotationProperty`	Intensive use of Dublin Core
Attribute $at_i \in T$	`qb:AttributeProperty`	Link to existing datasets such as DBPedia
OWA operator p	SPARQL 1.1 aggregation operators	Other extensions depend on the RDF repository
Index, Component and Indicator	`skos:Concept`	SKOS taxonomy (logical structure)

Example of the "Naive Requirement Quality Index" structure in RDF.

```
@prefix rdfindex: <http://purl.org/rdfindex/ontology/> .
@prefix rdfindex-rq: <http://purl.org/rdfindex/rq/resource/> .
@prefix rdfindex-rqont: <http://purl.org/rdfindex/rq/ontology/> .

rdfindex-rq:TheNaiveRequirementQualityIndex
  a rdfindex:Index;
   rdfs:label "The Naive Requirement Quality Index"@en;
   rdfindex:type rdfindex:Quantitative;
   rdfindex:aggregates [
     rdfindex:aggregation-operator rdfindex:OWA;
     rdfindex:part-of [
       rdfindex:dataset rdfindex-rq:Maintainability;
       rdfindex:weight 0.4];
     rdfindex:part-of [rdfindex:dataset rdfindex-rq:Usability;
       rdfindex:weight 0.6];
   ];
  #More meta-data properties...
   qb:structure  rdfindex-rq:TheNaiveRequirementQualityIndexDSD ; .

rdfindex-rq:TheNaiveRequirementQualityIndexDSD
  a qb:DataStructureDefinition;
  qb:component
  [qb:dimension rdfindex-rqont:ref-project],
  [qb:measure   rdfindex:value],
  [qb:attribute sdmx-attribute:unitMeasure];
  #More meta-data properties...
  .
```

SPARQL template for building aggregated observations.

```
SELECT (d_i ∈ D) [(sum(?w*?measure) as ?newvalue) | OWA(?measure)]
WHERE{
  q rdfindex:aggregates ?parts.
  ?parts rdfindex:part-of ?partof.
  ?partof rdfindex:dataset q_i .
  FILTER(?partof ∈ Q).
  ?observation rdf:type qb:Observation.
  ?part rdfindex:weight ?defaultw.
  OPTIONAL {?partof rdfindex:weight ?aggregationw.}.
  BIND (if( BOUND(?aggregationw), ?aggregationw, ?defaultw) AS ?w)
  ?observation m ?measure .
  ?observation ?dim ?dimRef.
  FILTER (?dim ∈ D).
}GROUP BY (d_i ∈ D)
```

Example of generated SPARQL query.

```
prefix rdfindex: <http://purl.org/rdfindex/ontology/>
SELECT ?dim0 ( sum(?w*?measure) as ?newvalue)
WHERE{
  rdfindex-rq:TheNaiveRequirementQualityIndex
    rdfindex:aggregates ?parts.
  ?parts rdfindex:part-of ?partof.
  ?partof rdfindex:dataset ?part .
  FILTER ((?part =rdfindex-rq:Maintainability) ||
  (?part =rdfindex-rq:Usability)).
  ?observation qb:dataSet ?part .
  ?part rdfindex:weight ?defaultw.
  OPTIONAL {?partof rdfindex:weight ?aggregationw.}.
  BIND (if( BOUND(?aggregationw), ?aggregationw, ?defaultw) AS ?w)
  ?observation rdfindex:value ?measure .
  ?observation rdfindex-rqont:ref-project ?dim0.
} GROUP BY ?dim0 ?dim1
```

Partial example of a populated observation for "The Naive Requirement Quality Index".

```
rdfindex-rq:o6808100851579
       a          qb:observation  ;
       qb:dataSet  rdfindex-rq:TheNaiveRequirementQualityIndex   ;
       rdfindex-rqont:ref-project  rdfindex-rq:P1;
       dcterms:date <http://reference.data.gov.uk/id/cd
gregorian-interval/2013-01-01T00:00:00/P1Y> ;
       ...
       #rdfs:{label,comment} {literal};
       #dcterms:{issued, date, contributor, author, publisher, identif
       # {resource, literal};
       rdfindex:md5-hash "002e1a2c78e41d7312ddd99e46bcbf41";
       sdmx-concept:obsStatus
               sdmx-code:obsStatus-E;
       rdfindex:value "0.61"^^xsd:double  .
```

On-going working examples

The Cloud Index

- Compilation of **Key Performance Indicators** for cloud services.
- Creation of a Cloud Index to measure **Quality of Service** of cloud providers.
- Funded by the RELATE-ITN (FP7-PEOPLE-2010-ITN, 264840) project and developed in the subproject **"Quality Management in Service-based Systems and Cloud Applications"**.
- Demo: `http://cloudindex-doc.herokuapp.com/`

RQA of The Reuse Company

- Compilation of several metrics (32) .
- Creation of the Requiremens Quality Index to measure the CCC of a specification.
- Expose an OSLC-based interface to access metric values, etc.
- Funded by the Carlos III University of Madrid and The Reuse Company within the Crystal project.

On-going working examples

The Cloud Index

- Compilation of **Key Performance Indicators** for cloud services.
- Creation of a Cloud Index to measure **Quality of Service** of cloud providers.
- Funded by the RELATE-ITN (FP7-PEOPLE-2010-ITN, 264840) project and developed in the subproject **“Quality Management in Service-based Systems and Cloud Applications”**.
- Demo: `http://cloudindex-doc.herokuapp.com/`

RQA of The Reuse Company

- Compilation of several metrics (32) .
- Creation of the Requiremens Quality Index to measure the CCC of a specification.
- Expose an OSLC-based interface to access metric values, etc.
- Funded by the Carlos III University of Madrid and The Reuse Company within the Crystal project.

Demo of the RDFIndex through an OSLC-based interface...

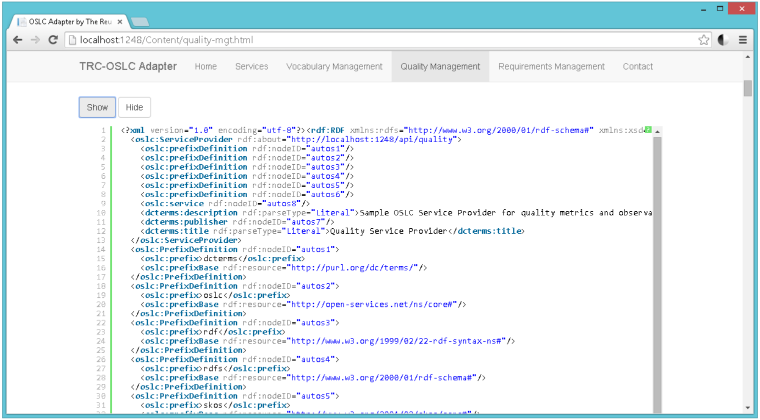


Figure: Description of the OSLC-based service provider.

Demo of the RDFIndex through an OSLC-based interface...

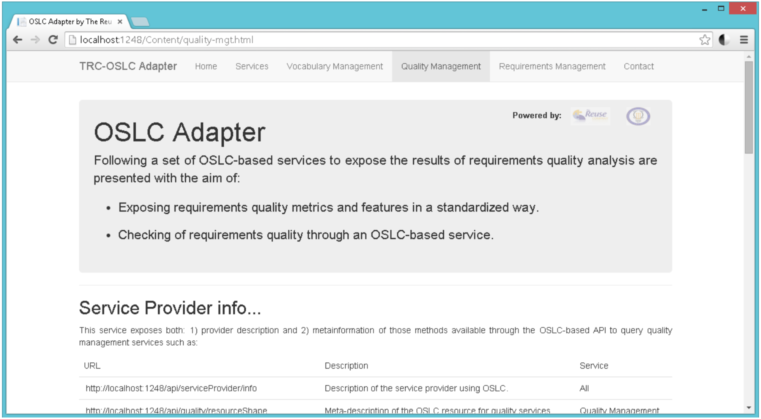


Figure: Description of the OSLC-based service provider.

Demo of the RDFIndex through an OSLC-based interface...

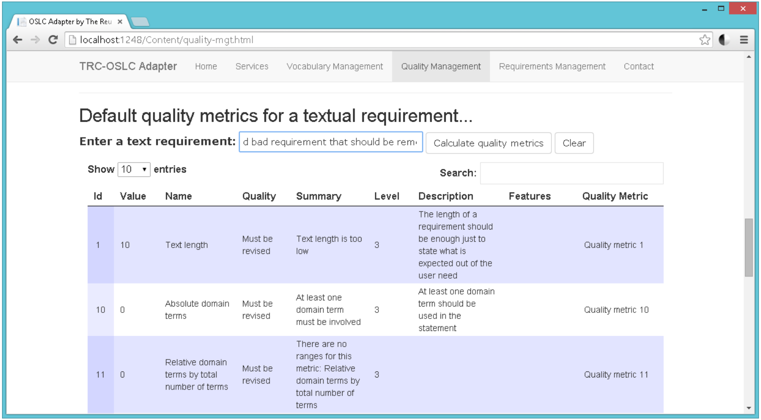


Figure: Calculating RQA quality metrics and presenting as HTML.

Demo of the RDFIndex through an OSLC-based interface...

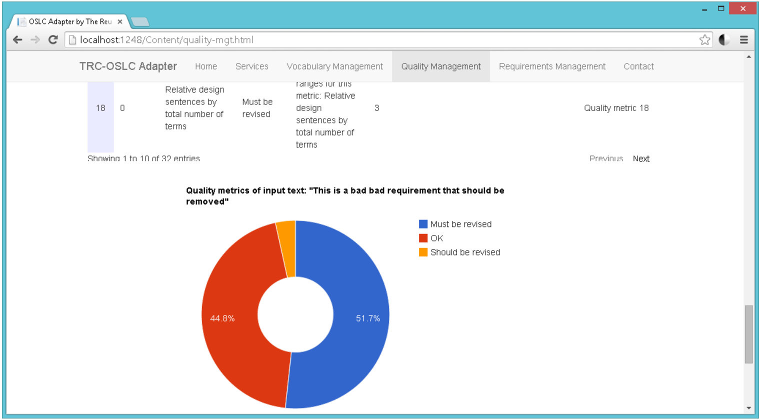


Figure: Calculating RQA quality metrics and graphical view with D3.js.

Demo of the RDFIndex through an OSLC-based interface...

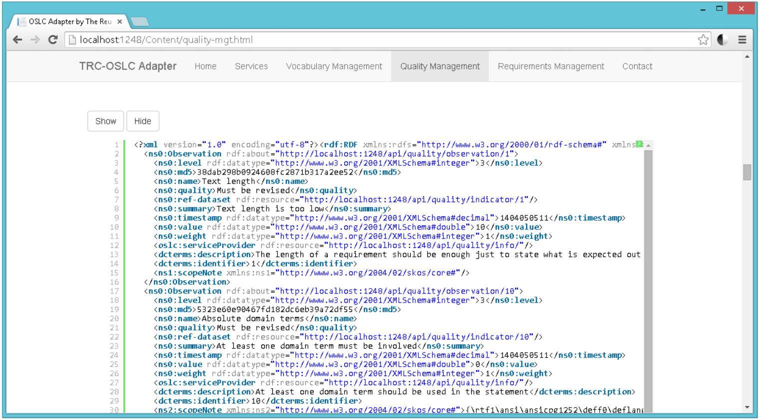


Figure: Calculating RQA quality metrics and RDF-based representation.

Advantages

Data Sources

Data management applying the Linked Data principles.

Structure

Automatic validation and verification of the structure and metadata of quantitative indexes.

Computation process

A native approach using SPARQL queries that can help to boost transparency.

Documentation

Implicit multilingual and multicultural support.

Advantages

Data Sources

Data management applying the Linked Data principles.

Structure

Automatic validation and verification of the structure and metadata of quantitative indexes.

Computation process

A native approach using SPARQL queries that can help to boost transparency.

Documentation

Implicit multilingual and multicultural support.

Advantages

Data Sources

Data management applying the Linked Data principles.

Structure

Automatic validation and verification of the structure and metadata of quantitative indexes.

Computation process

A native approach using SPARQL queries that can help to boost transparency.

Documentation

Implicit multilingual and multicultural support.

Advantages

Data Sources

Data management applying the Linked Data principles.

Structure

Automatic validation and verification of the structure and metadata of quantitative indexes.

Computation process

A native approach using SPARQL queries that can help to boost transparency.

Documentation

Implicit multilingual and multicultural support.

Detailed advantages (I)

Feature	Main advantages
Data sources	• Common and shared data model, RDF. • Description of data providers (provenance and trust). • A formal query language to query data, SPARQL. • Use of Internet protocols, HTTP. • Data enrichment and validation (domain and range). • Unique identification of entities, concepts, etc. through (HTTP) URIs. • Possibility of publishing new data under the aforementioned characteristics. • Standardization and integration of data sources.
Structure	• Meta-description of index structure (validation). • Re-use of existing semantic web vocabularies. • Re-use of existing datasets to enrich meta-data. • Context-aware definitions. • Underlying logic formalism. • Orthogonal and flexible.

Detailed advantages (II)

Computation process	• Meta-description of datasets aggregation. • Validation of composed datasets. • OWA operators support. • Direct translation to SPARQL queries.
Documentation	• Multilingual support to describe datasets, etc. • Easy generation with existing tools.
Cross-Domain Features	• Separation of concerns and responsibilities: data and meta-data (structure and computation). • Standardization (put in action specs from organisms such as W3C). • Declarative and adaptive approach. • Non-vendor lock-in (format, access and computation). • Integration, Interoperability and Transparency. • Align to existing trend (data management: quality and filtering). • Easy integration with third-party services such as visualization. • Contribution to the Web of Data.

Restrictions

SPARQL 1.1 support

1. Some built-in functions are not standardized:
   - Example: `z-score` employs standard deviation.
   - It requires built-in function: `sqrt` (only available in some SPARQL implementations).
2. Ranking of values is not obvious:
   - `GROUP_CONCAT`
   - Check a value against all the other values.
   - ...neither solution is efficient.

Limitations of the RDFIndex expressivity

The working examples show its applicability but some lack of terms/relationships should be expected.

Restrictions

SPARQL 1.1 support

1. Some built-in functions are not standardized:
   - Example: `z-score` employs standard deviation.
   - It requires built-in function: `sqrt` (only available in some SPARQL implementations).
2. Ranking of values is not obvious:
   - `GROUP_CONCAT`
   - Check a value against all the other values.
   - ...neither solution is efficient.

Limitations of the RDFIndex expressivity

The working examples show its applicability but some lack of terms/relationships should be expected.

Conclusions

1. The use of quantitative indexes is a common practice to compile and rank indicators for making decisions.
2. Existing indexes are completely opaque (PDF, HTML, etc.)
3. It is necessary to **improve** the **access to data/metadata** and to **boost transparency**.
4. The **RDFIndex** vocabulary seeks for providing the **appropriate building blocks** to **represent** and **compute** indexes.
5. However:
    - SPARQL limitations avoid a fully development of a "pedantic" Linked Data approach.
    - The vocabulary likely does not cover all required elements to represent any index.

Future Work

1. Extend the RDFIndex vocabulary to enhance the expressivity (more numerical values, metric units, etc.).
2. Standardize the OSLC interface.
3. Externalize the computation with a hybrid approach R+SPARQL.
4. Parallelization of the computation process.
5. Event-driven and continuos calculation of requirements quality.


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EKHMET
GRAZIE
MEHRBANI
BOLZİN
MERCI
BİYAN
SHUKRIA
TINGKI
GOZAIMASHITA
EFCHARISTO
JUSPAXAR
KOMAPSUMNIDA
MAAKE
PALDIES

Roster...

- Dr. Jose María Alvarez-Rodríguez
- Knowledge Reuse Group
- Carlos III University of Madrid, Spain
- E-mail: josemaria.alvarez@uc3m.es
- WWW: http://www.josemalvarez.es

- Prof. Dr. Juan Llorens
- Knowledge Reuse Group
- Carlos III University of Madrid, Spain
- E-mail: llorens@uc3m.es
- WWW: https://www.linkedin.com/pub/juan-llorens/b/857/632

Roster (II)...

- Ms. Manuela Alejandres
- Knowledge Reuse Group
- Carlos III University of Madrid, Spain
- E-mail: manoli@kr.inf.uc3m.es

- Mr. Jose Fuentes
- The Reuse Company (TRC), Spain
- E-mail: jose.fuentes@reusecompany.com
- WWW: https://reusecompany.com

Z-score normalization using SPARQL...

```
prefix afn: <http://jena.hpl.hp.com/ARQ/function#>
SELECT ( (?measure-?mean)/?stddev as ?zscore)
WHERE{
 ...
 ?observation rdfindex:value ?measure
 {
  SELECT ?mean (afn:sqrt((SUM((?measure-?mean)*
 (?measure-?mean))/?count)) as ?stddev)
  WHERE{
   ?observation rdfindex:value ?measure
   {
    SELECT (COUNT(?measure) as ?count) (AVG(?measure) as ?mean)
    WHERE {
     ?observation rdfindex:value ?measure
     }GROUP BY ?count ?mean LIMIT 1

   }
  }GROUP BY ?mean ?count LIMIT 1
 }
}
```

T. Bosch, R. Cyganiak, A. Gregory, and J. Wackerow.
DDI-RDF discovery vocabulary. a metadata vocabulary for documenting research and survey data.
In 6th Workshop on Linked Data on the Web (LDOW2013), 2013.

R. Cyganiak and D. Reynolds.
The RDF Data Cube Vocabulary.
Working Draft, W3C, 2013.
http://www.w3.org/TR/vocab-data-cube/.

I. Ermilov, M. Martin, J. Lehmann, and S. Auer.
Linked Open Data Statistics: Collection and Exploitation.
In KESW, pages 242–249, 2013.

J. D. Fernández, M. A. Martínez-Prieto, and C. Gutiérrez.
Publishing open statistical data: the Spanish census.
In DG.O, pages 20–25, 2011.

M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers.
SCOVO: Using Statistics on the Web of Data.
In L. A. et al., editor, The Semantic Web: Research and Applications, volume 5554 of Lecture Notes in Computer Science, pages 708–722. Springer Berlin Heidelberg, 2009.

J. M. A. Rodríguez, J. Clement, J. E. L. Gayo, H. Farhan, and P. Ordoñez De Pablos.
Publishing Statistical Data following the Linked Open Data Principles: The Web Index Project, pages 199–226.
IGI Global, 2013.

P. E. R. Salas, F. M. D. Mota, K. K. Breitman, M. A. Casanova, M. Martin, and S. Auer.
Publishing Statistical Data on the Web.
Int. J. Semantic Computing, 6(4):373–388, 2012.

B. Zapilko and B. Mathiak.
Defining and Executing Assessment Tests on Linked Data for Statistical Analysis.
In COLD, 2011.

Towards a semantic-based representation and computation of quantitative indexes for quality management of requirements. The RDFIndex approach.

Jose María Álvarez-Rodríguez (speaker)
and
{Juan Llorens, Manuela Alejandres and
Jose Miguel Fuentes}

INCOSE 2014 | 24th Annual INCOSE International Symposium
Las Vegas, USA