



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

»» Metainformação comum ao serviço da interoperabilidade estatística

O papel das ontologias e da Web semântica

Sérgio Bacelar
(INE)

 (Lisboa, 8 de Novembro 2011)



Desafios actuais do Institutos de Estatística

- Procura de estatísticas de alta qualidade
- Restrições
 - Cortes orçamentais
 - Competição intensificada
 - Consequências da globalização
 - Diminuição da carga sobre os respondentes
- Resposta **comum** da comunidade estatística
 - Reformatação dos processos de negócio
 - Standardização
 - Interoperabilidade

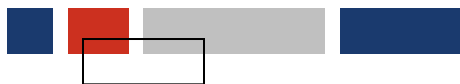




Desenvolvimento da interoperabilidade

- Adopção de standards com sucesso
 - Desenvolvimento ao longo do tempo
 - Maturação por fases
- Ponto de partida típico
 - Necessidade de partilha de dados
 - Utilização de standards
 - Partilha de instrumentos
 - Intensificação da cooperação





Três pilares para o incremento da interoperabilidade

Aspectos organizacionais

- Semelhanças na base legal do negócio
- Globalização, comparabilidade, harmonização de produtos

Aspectos semânticos

- Classificações e definições comuns
- Necessidade de harmonizar a informação e os outputs

Questões técnicas

- Desenvolvimento e utilização comum de instrumentos, standards técnicos e tecnologias





Interoperabilidade do ponto de vista da comunidade estatística

- Forma como gerimos o nosso negócio (processos de negócio)
- O que produzimos (dados e análises estatísticas)
- Métodos e instrumentos de apoio ao negócio





Iniciativas (1)

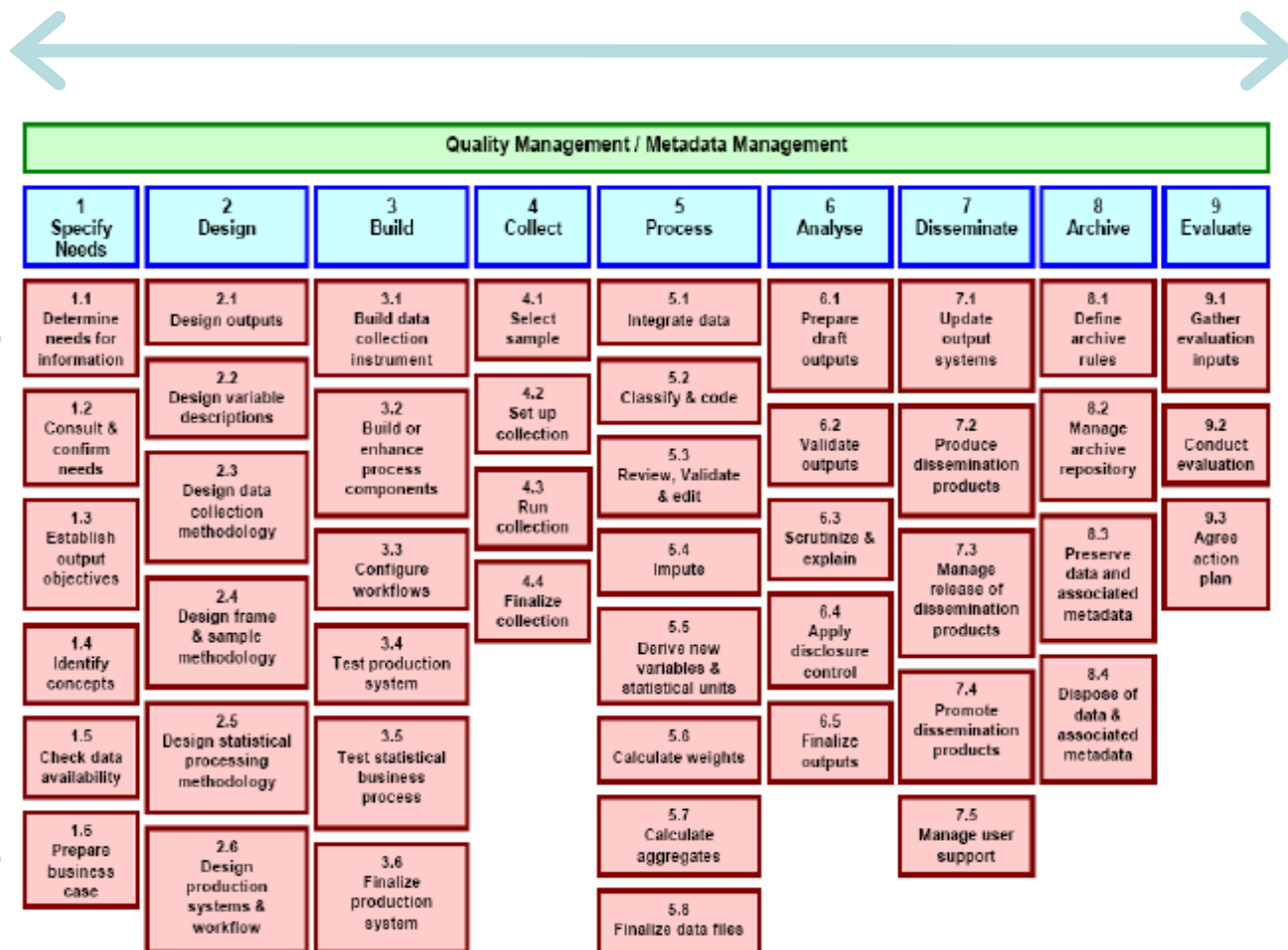
- Generic Statistical Business Process Model (UNECE)
 - Modelo que descreve os processos de negócio dos Institutos de Estatística
 - Harmonização da produção estatística
- Necessidade de que os fluxos de dados sejam consistentes com um modelo conceptual
 - Standardização de dados e de metainformação

GSBPM - Estrutura do modelo

Processo

Fases

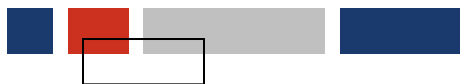
Sub-processos





Iniciativas (2)

- SDMX (Statistical Data and Metadata eXchange)
 - Decisão de usar o SDMX em todas as comunicações de dados (*Comissão Estatística das Nações Unidas*)
 - Interoperabilidade sintáctica como um pré-requisito para a interoperabilidade semântica
 - *Information Model e Metadata Common Vocabulary (MCV)*
- DDI (*Data Documentation Initiative*)
- GSIM (*Generic Statistical Information Model*)



Metainformação e transmissão da informação estatística

- Transmissão de:
 - Informação estatística agregada...
 - ...e da correspondente metainfomação estrutural e de referência
 - Standard internacional: *Statistical Data and Metadata EXchange* (SDMX).
 - Conjunto de vocabulários controlados de uso comum, estruturados num modelo de dados e de metadados.





Interoperabilidade semântica: desafios

- Ausência de um vocabulário comum de metainformação, harmonizado e passível de referenciação computacional.
- ESSnet on SDMX – Rede de excelência do Eurostat (<http://sdmxxessnet.ine.pt>)
- Construção duma ontologia adaptada à web semântica (*MCV Ontology: Workpackage 2 – WP2*)





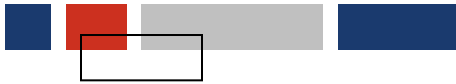
Ontologias - justificação

- Ontologia
 - é um modelo organizado de conhecimento num dado domínio. Componentes: classes, atributos, relações e instâncias
- Objectivo: dar significado à informação
 - Ambiguidade do significado de determinados termos
 - Resultados de pesquisa de grande dimensão, sem ordem lógica



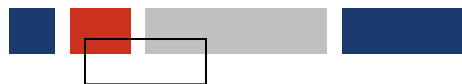
Metadata Common Vocabulary ontology

Objectivo, domínio e âmbito



- Objectivo
 - Construir uma ontologia de forma a criar condições de desenvolvimento de uma proposta de versão estruturada de um vocabulário comum de metainformação (MCV)
- Domínio
 - O quadro de referência é o domínio da **metainformação estatística**
- Âmbito
 - Utilização da metainformação no contexto da **iniciativa SDMX**





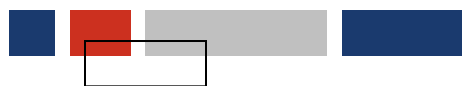
Metodologia de trabalho

Classificação de conceitos

Três categorias:

1. **Conceitos gerais** – conceitos fundamentais e genéricos
2. **Informação estatística** – contextualizam a metainformação estatística (domínio da ontologia)
3. **Actividade estatística** – representam o campo onde o domínio é aplicado





Metodologia de trabalho

Fontes

Metainformação disponível no MCV

- Natureza heterogénea e interdisciplinar

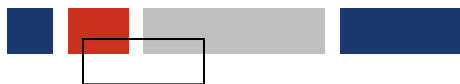
Quadros conceptuais diferentes

Conceitos provenientes de fontes diferentes

Outras fontes

- SDMX User Guide
- SDMX Information Model (SDMX-IM)
- Outros documentos produzidos em organizações estatísticas





Resultados

Seleção de conceitos

Conceitos incluídos na ontologia

Conceitos	n
Definição alterada	172
Novo termo	33
Termo alterado	3
Termo/definição alterado	2
Nova definição	4
Nova fonte	1
Definição não alterada	38
TOTAL	253



Resultados

Representação de conceitos

- **Exemplo:**

- **Statistical macrodata**

- Definição (MCV): An observation data gained by a purposeful aggregation of statistical microdata conforming to statistical methodology.
 - Definição (WP2): statistical data obtained by aggregation of statistical microdata.



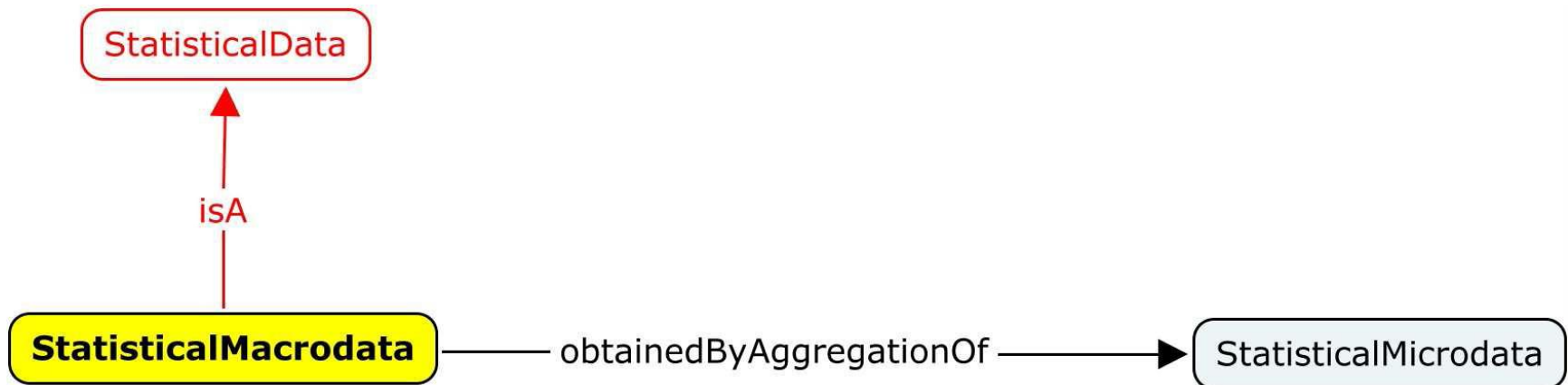


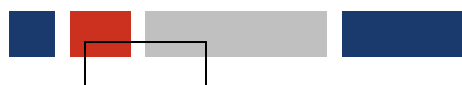
Resultados

Representação de conceitos

Definições em **linguagem semi-formal** (CMap Tools)

De acordo com convenções metodológicas e gráficas.





Resultados

Representação de conceitos

Definições em linguagem formal (OWL) com Protégé

Superclasses 

● **StatisticalData**

● **isObtainedByAggregationOf** **some** StatisticalMicrodata

Inherited anonymous classes

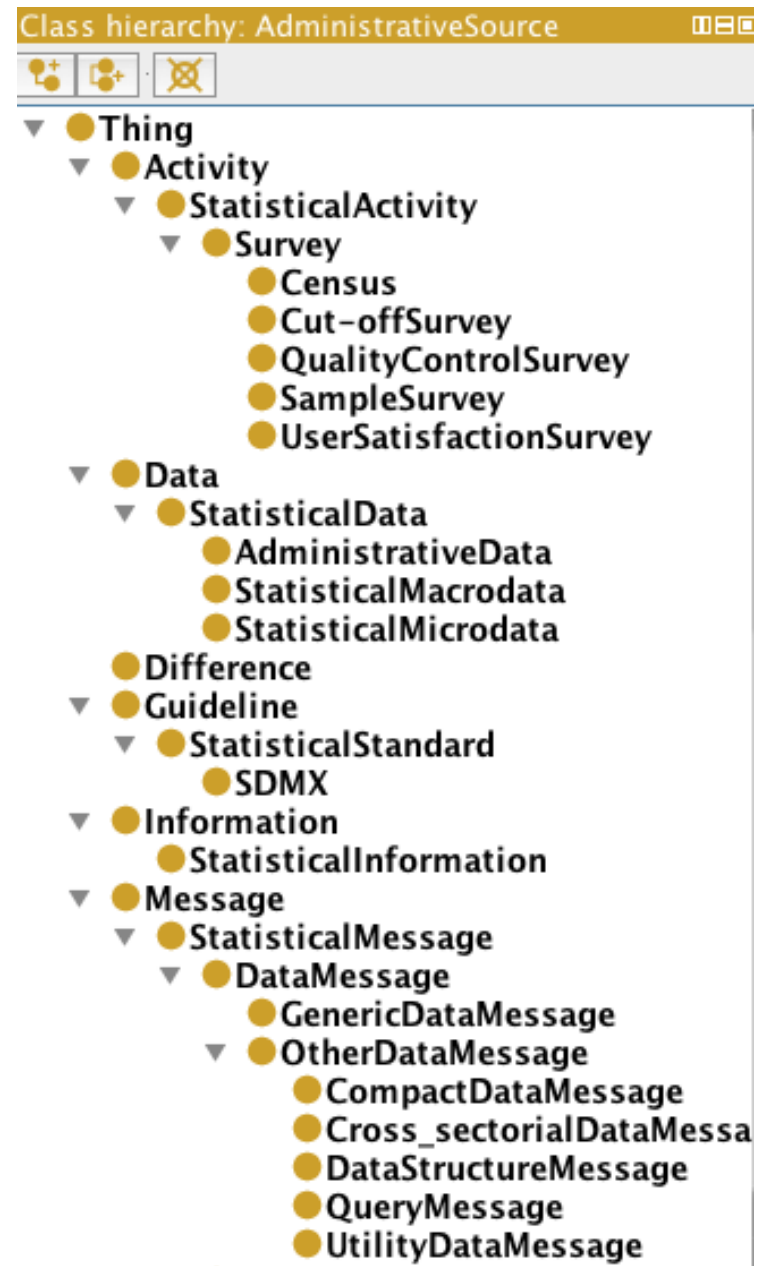
● **isObtainedThrough** **some** StatisticalActivity

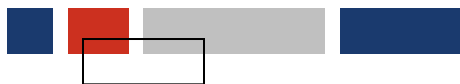


Resultados

Representação de conceitos

- Classes e hierarquia de classes em OWL
 - Relações do tipo “kind-of”
 - Parsimonia na definição do nº de classes → clareza da ontologia





Definição de **propriedades de objecto**

Expressão de relações entre conceitos (classes)

- Para relacionar o conceito de <Statistical Message> com os conceitos <Statistical Standard>, <Data Set> e <Metadata Set> usaram-se duas **Propriedades de Objecto**:
 - **accordingTo** e **contains**.

- Com estas obtêm-se as seguintes **expressões**:

<Statistical Message> accordingTo **some** <StatisticalStandard>;
<Statistical Message> contains **some** <DataSet>;
<Statistical Message> contains **some** <MetadataSet>.

“**Some**” representa um quantificador de existência;



Hierarquia das propriedades de objectos



Statistical Message



Superclasses 

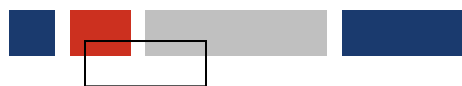
- **Message**

- **accordingTo *some* StatisticalStandard**

- **contains *some* DataSet**

- **contains *some* MetadataSet**





Resultados

Representação de conceitos

Propriedades que designam atributos de conceitos

- “QualityDimension” é um atributo da classe <Product>
- Usou-se uma propriedade de objecto do tipo “has Attribute”, tal como “hasQualityDimension” e criou-se a super-classe [hasQualityDimension> some <QualityDimension>].

Superclasses 

 **hasQualityDimension** **some** QualityDimension

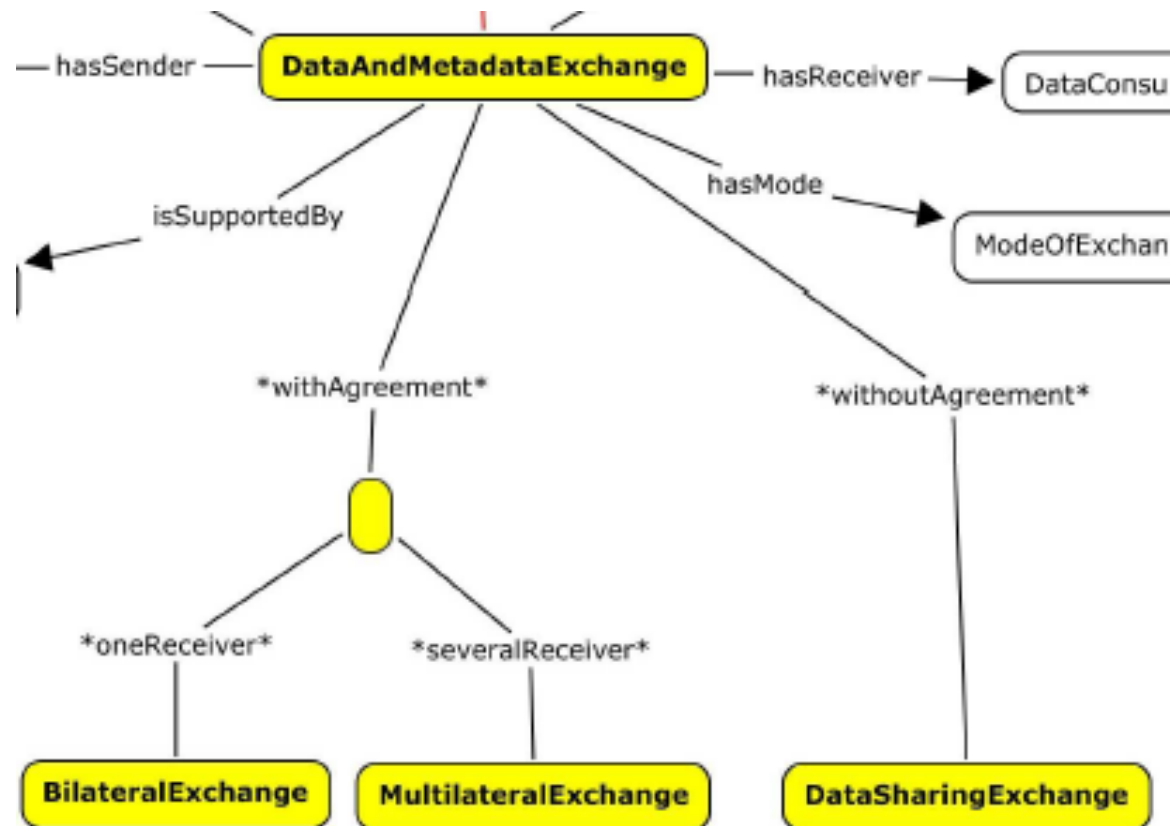


Resultados

Representação dos conceitos

Caso especial:

- Para distinguir sub-classes que pertencem à mesma classe, usou-se a classe '**Difference**' e os **indivíduos** correspondentes para representar a *diferença específica*.



Resultados

Representação dos conceitos

Troca partilhada

Superclasses +

- DataAndMetadataExchange
- differences value withoutAgreement

Troca bilateral

Superclasses +

- DataAndMetadataExchange
- differences value oneReceiver
- differences value withAgreement

Troca multilateral

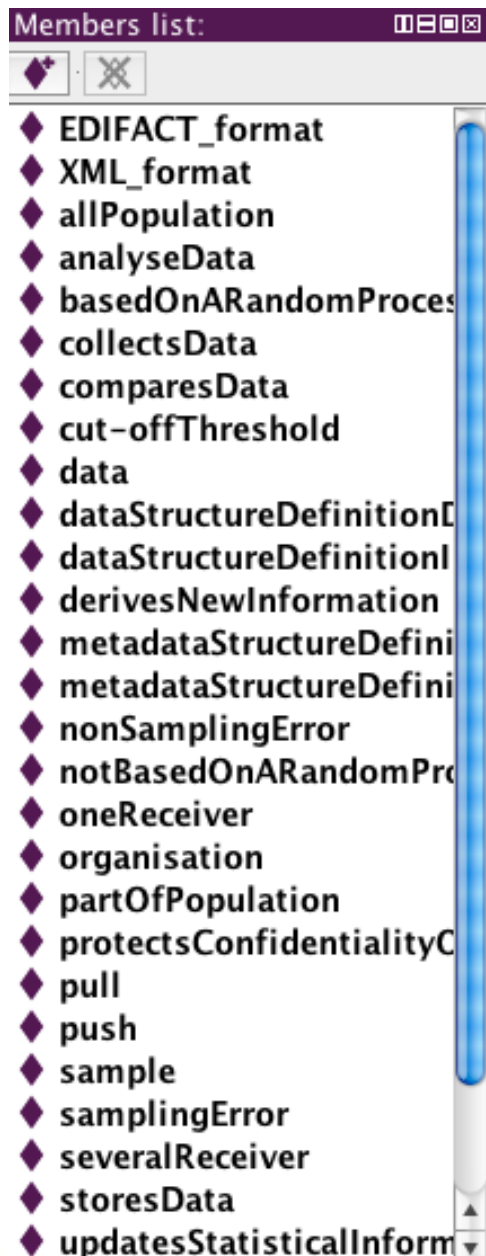
Superclasses +

- DataAndMetadataExchange
- differences value severalReceiver
- differences value withAgreement



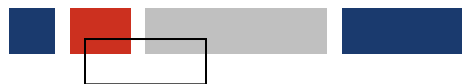
Indivíduos

Diferença



Erro não amostral





Resultados

Representação de conceitos

Nalguns casos, utilizaram-se propriedades inversas como:

“uses” e “isUsed”

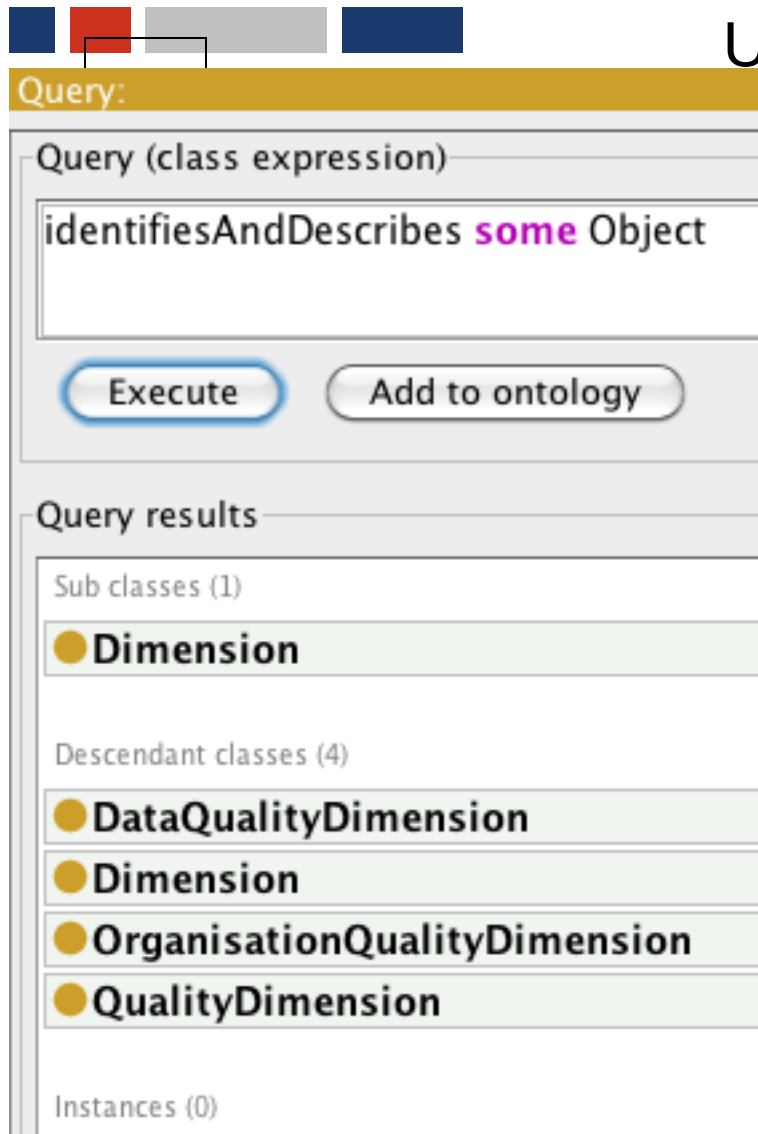
“represents” e “isRepresentedBy”

Uses



Interrogar a ontologia

Utilização do DL-Query no Protégé



The screenshot shows the Protégé DL-Query interface. At the top, there is a 'Query:' label. Below it, a text box contains the query 'identifiesAndDescribes some Object'. To the right of the text box are two buttons: 'Execute' and 'Add to ontology'. Below the query box is a section titled 'Query results'. Under this section, there are three categories: 'Sub classes (1)', 'Descendant classes (4)', and 'Instances (0)'. Under 'Sub classes (1)', there is one result: 'Dimension'. Under 'Descendant classes (4)', there are four results: 'DataQualityDimension', 'Dimension', 'OrganisationQualityDimension', and 'QualityDimension'. Each result is preceded by a yellow circle icon.

Query:

Query (class expression)

identifiesAndDescribes some Object

Execute Add to ontology

Query results

Sub classes (1)

- Dimension

Descendant classes (4)

- DataQualityDimension
- Dimension
- OrganisationQualityDimension
- QualityDimension

Instances (0)

Em linguagem
natural

Devolver todos os
conceitos que
identificam e
descrevem algum
Objecto.



Interrogar a ontologia

Utilização do DL-Query no Protégé



Query:

Query (class expression)

identifiesAndDescribes some Object and (differences value data)

Execute Add to ontology

Query results

Sub classes (1)

● DataQualityDimension

Descendant classes (1)

● DataQualityDimension

Instances (0)

Em linguagem
natural

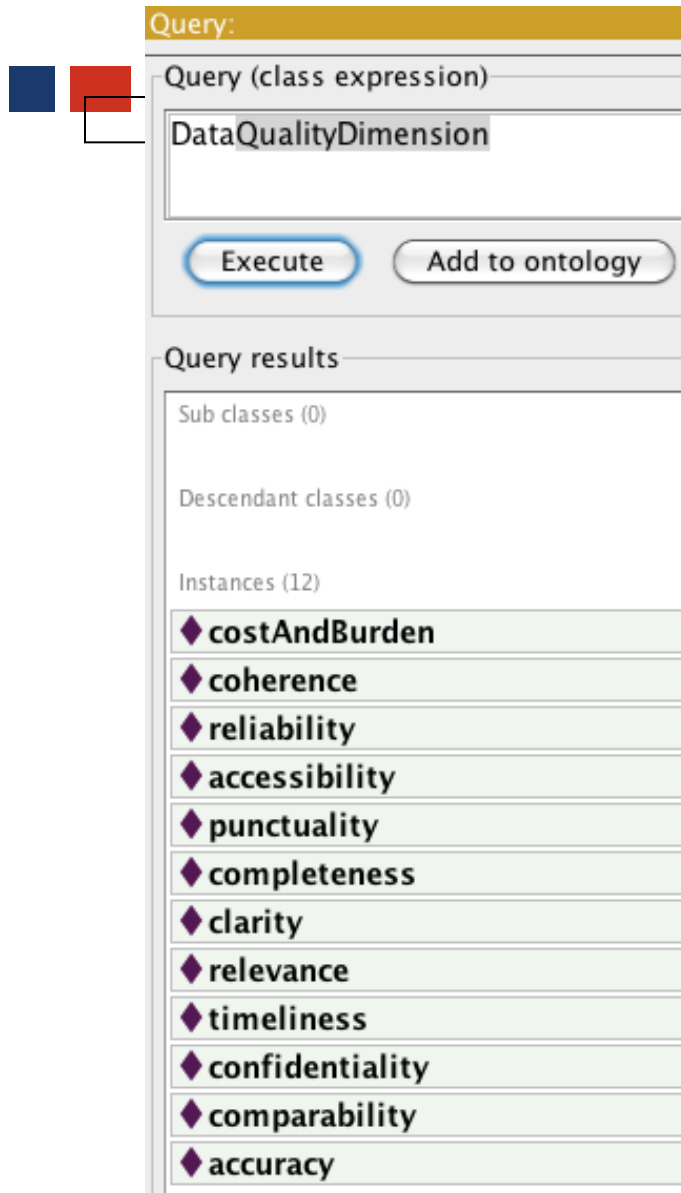
Devolver todos os conceitos que identificam e descrevem algum **Objecto** e que se distinguem doutros conceitos pelo facto de terem como objecto os dados.



Interrogar a ontologia
Utilização do DL-Query no Protégé

Em linguagem natural

Devolver todas as
instâncias de ***Data
Quality Dimension***.



The screenshot shows the Protégé DL-Query interface. At the top, there is a yellow header labeled "Query:". Below it, a text box contains the query "DataQualityDimension". To the right of the text box are two buttons: "Execute" (highlighted with a blue border) and "Add to ontology". Below the query box, there is a section titled "Query results". Under this section, there are three sub-sections: "Sub classes (0)", "Descendant classes (0)", and "Instances (12)". The "Instances (12)" section is expanded, showing a list of 12 instances, each preceded by a purple diamond icon. The instances are: costAndBurden, coherence, reliability, accessibility, punctuality, completeness, clarity, relevance, timeliness, confidentiality, comparability, and accuracy.

Instances (12)
◆ costAndBurden
◆ coherence
◆ reliability
◆ accessibility
◆ punctuality
◆ completeness
◆ clarity
◆ relevance
◆ timeliness
◆ confidentiality
◆ comparability
◆ accuracy



Software utilizado



- Mapas conceptuais
 - IHMC CMapTools (<http://cmap.ihmc.us/>)
- Editor de ontologias
 - Protégé (<http://protege.stanford.edu/>)

Conclusões

- Melhoria da
 - Organização e estrutura do MCV
 - Percepção das relações entre conceitos
 - Qualidade das definições
 - Promoção da utilização do SDMX
- Representação do domínio de conhecimento do MCV na linguagem OWL
 - Linguagem standard para representação do conhecimento
 - Aplicações baseadas na Ontologia
 - Possibilidade de utilização na Web Semântica

