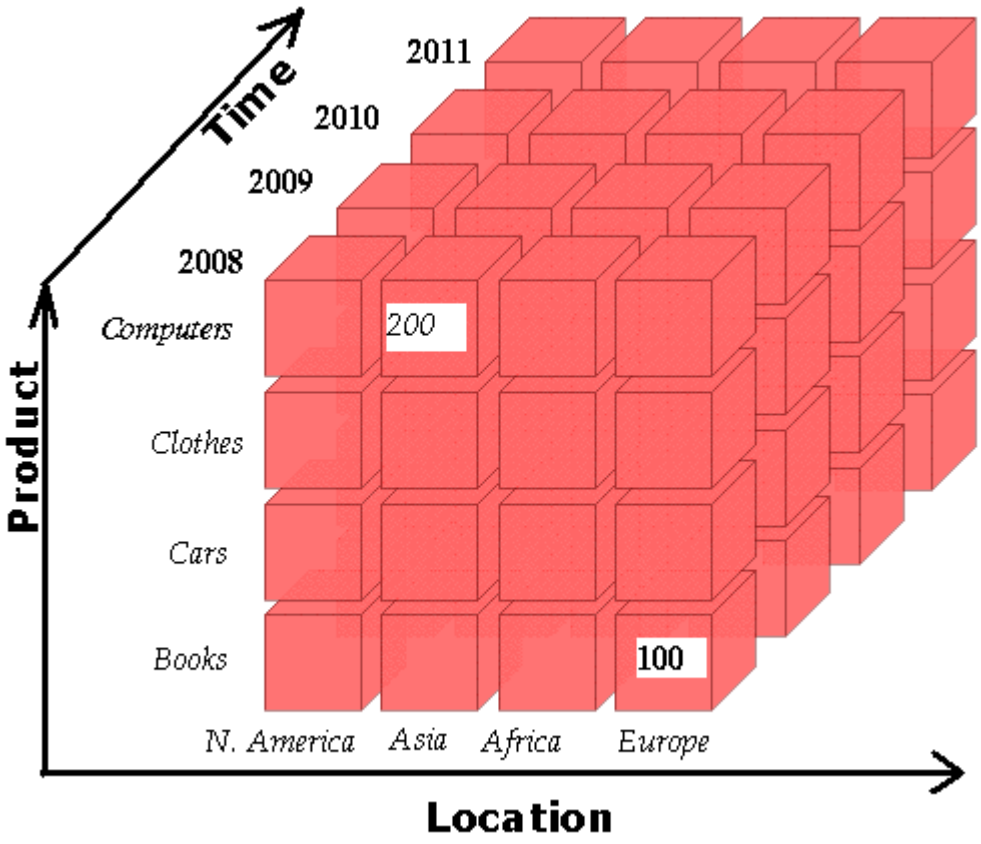
Data cubes and NOR_2O

Esther Lozano, Boris Villazón-Terrazas, Oscar Corcho
Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo sn, 28660 Boadilla del Monte, Madrid
http://www.oeg-upm.net
elozano@fi.upm.es
Phone: (+34) 913366605, Fax: (+34) 913524819

Data cubes

- A cube can be considered a multidimensional extension of 2-D tables (as in geometry a cube is a three-dimensional extension of a square)
- The term hypercube is sometimes used, especially for data with more than three dimensions
- Each dimension represents some attribute in the database (such as sales, profits, expenses…)
- The cells in the data cube represent the measure of interest. For example, they could contain a count for the number of times that attribute combination occurs in the database, or the minimum, maximum, sum or average value of some attribute
- Queries are performed on the cube to retrieve decision support information

Data cubes

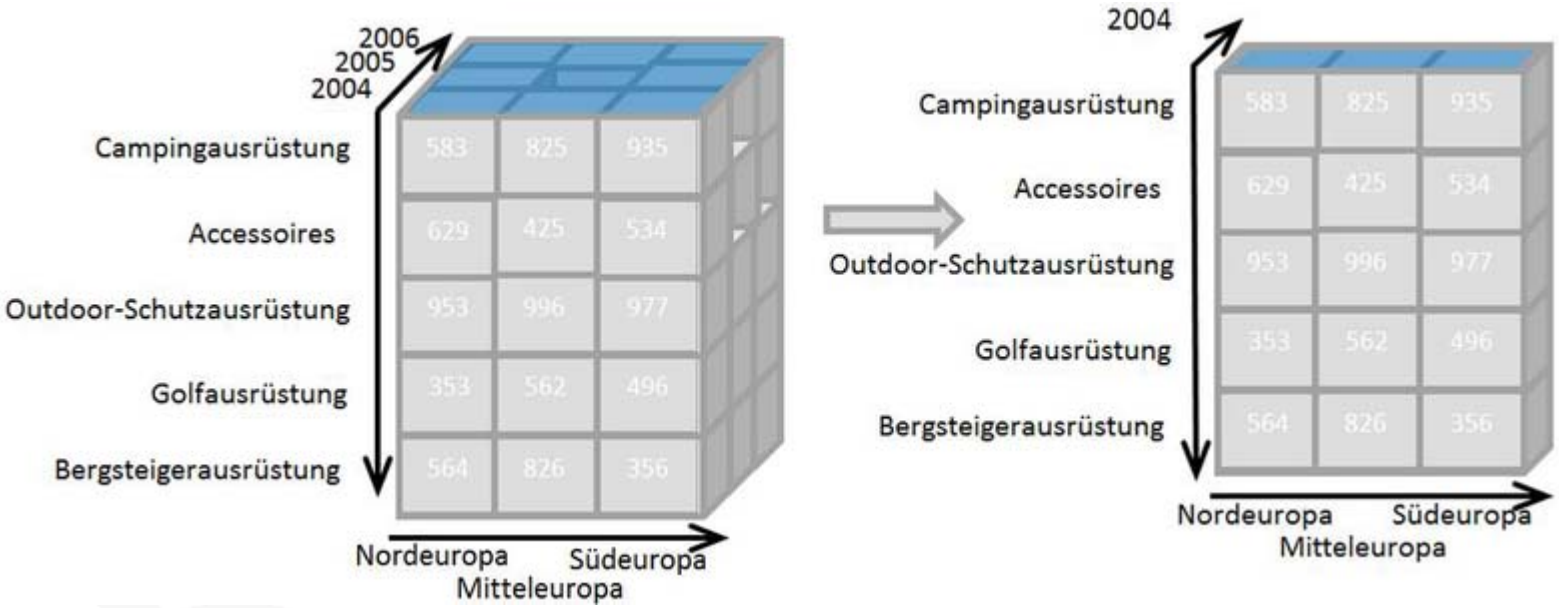

Time
2011
2010
2009
2008
Product
Computers
200
Clothes
Cars
Books
100
N. America
Asia
Africa
Europe
Location

OLAP Cube

- OLAP stands for Online Analytical Processing
- Computer-based technique for analyzing business data in the search for business intelligence
- The elements of a dimension can be organized as a hierarchy, where a parent member summarizes its children
- Parent elements can further be aggregated as the children of another parent
- Different operations to facilitate analysis, aligning the data content with a familiar visualization

Common OLAP Operations

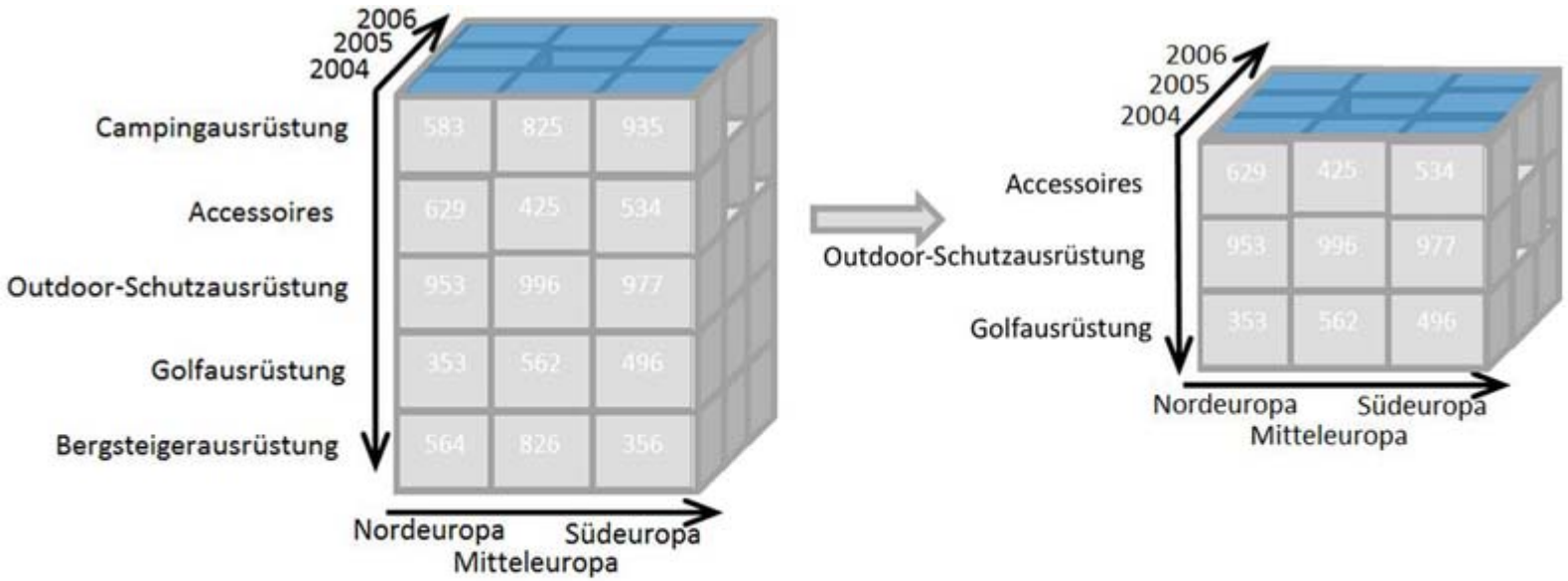
- *Slice* is the act of picking a rectangular subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension



http://en.wikipedia.org/wiki/OLAP_cube

Common OLAP Operations

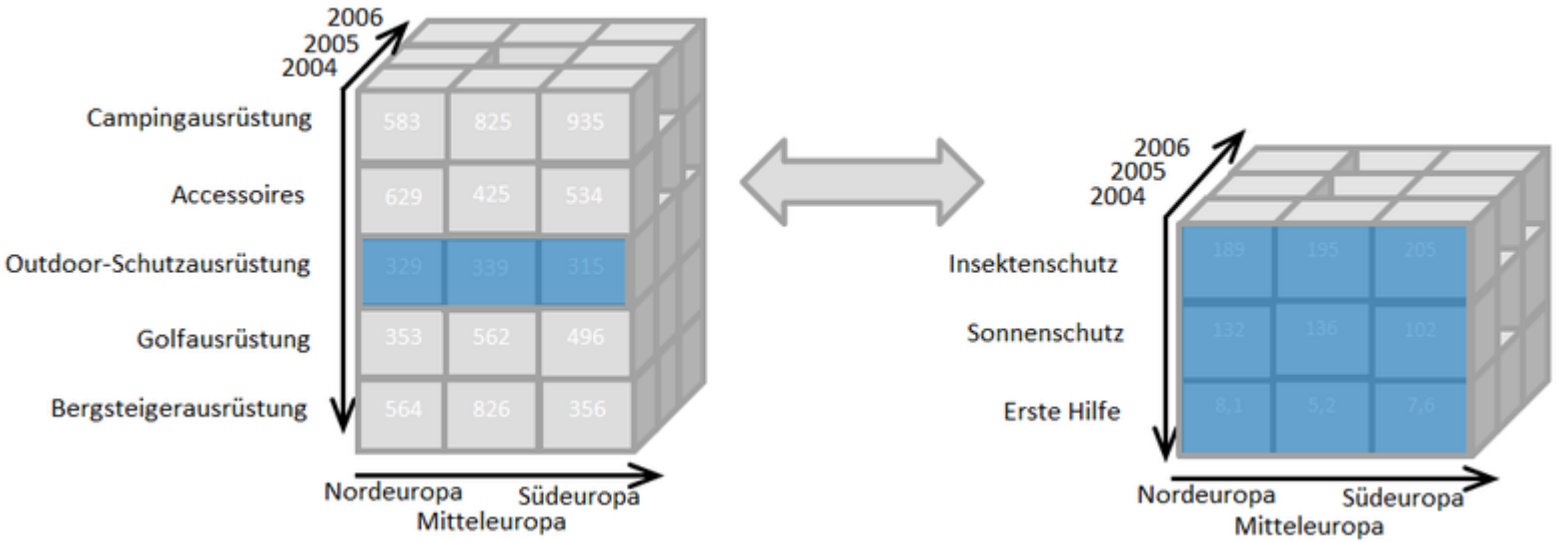
- *Dice* produces a subcube by allowing the analyst to pick specific values of multiple dimensions



http://en.wikipedia.org/wiki/OLAP_cube

Common OLAP Operations

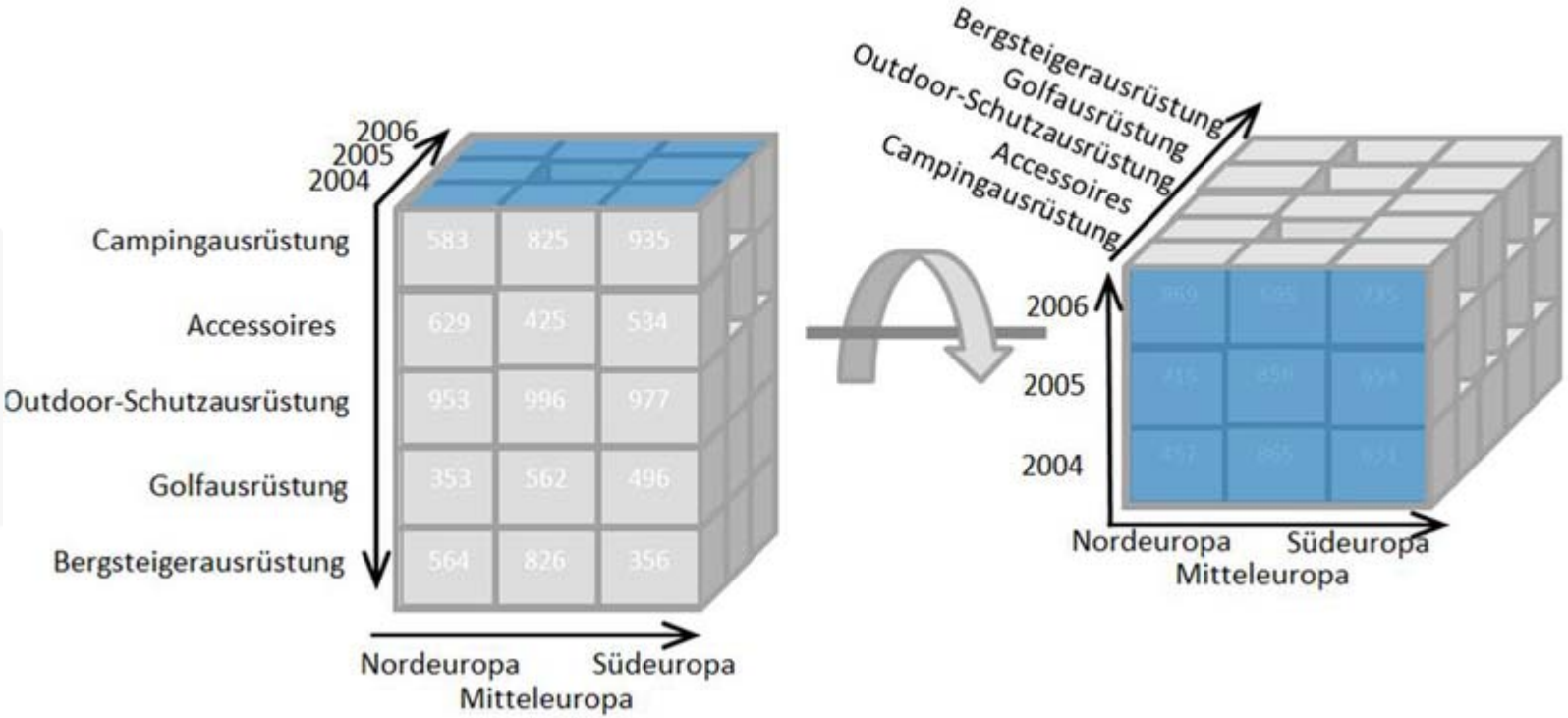
- *Drill Down/Up* allows the user to navigate among levels of data ranging from the most summarized (up) to the most detailed (down)



http://en.wikipedia.org/wiki/OLAP_cube

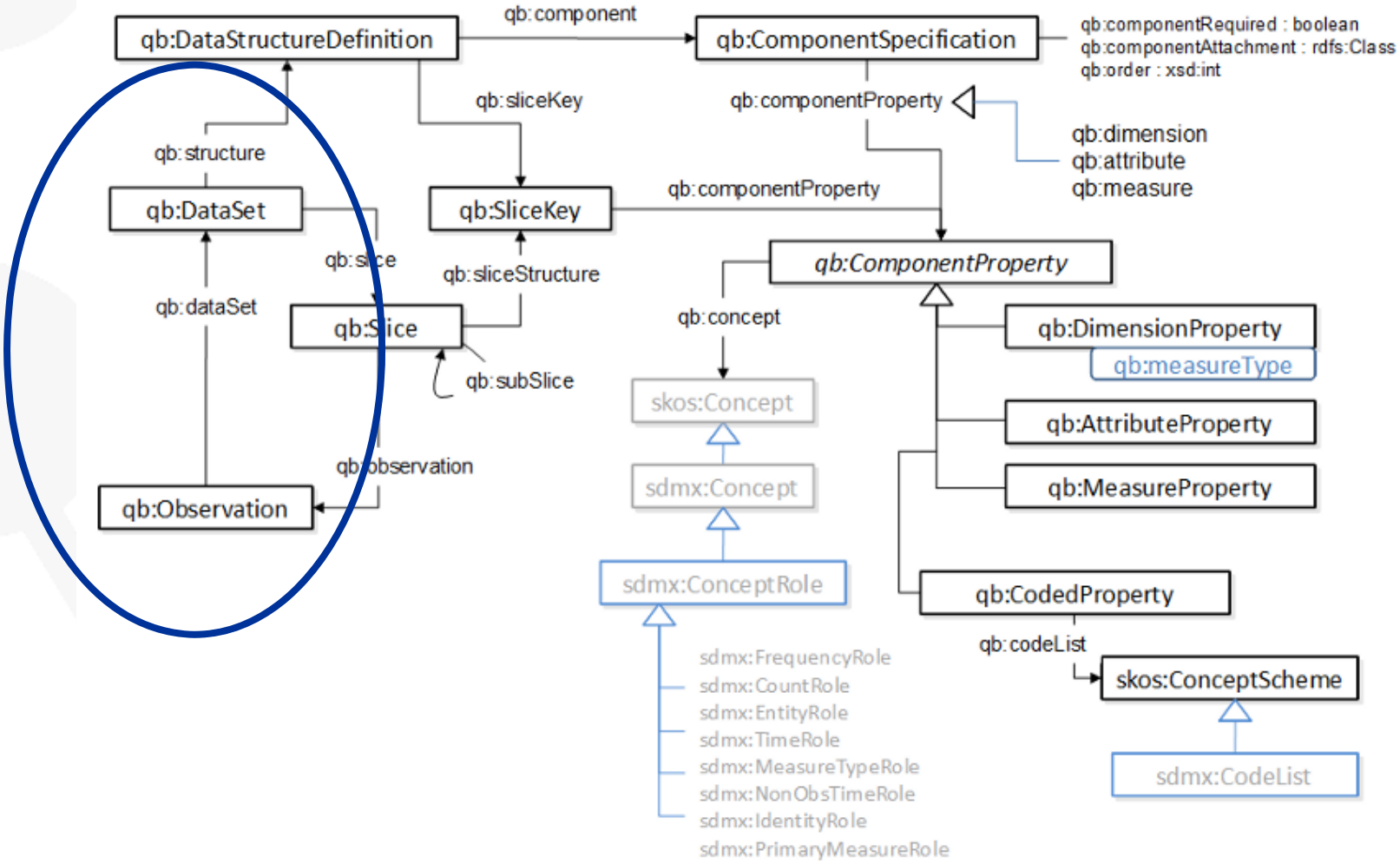
Common OLAP Operations

- *Pivot* allows an analyst to rotate the cube in space to see its various faces



http://en.wikipedia.org/wiki/OLAP_cube

- Vocabulary


qb:DataStructureDefinition
qb:component
qb:ComponentSpecification
qb:componentRequired : boolean
qb:componentAttachment : rdfs:Class
qb:order : xsd:int
qb:sliceKey
qb:componentProperty
qb:dimension
qb:attribute
qb:measure
qb:structure
qb:DataSet
qb:SliceKey
qb:componentProperty
qb:slice
qb:sliceStructure
qb:ComponentProperty
qb:dataSet
qb:Slice
qb:concept
qb:DimensionProperty
qb:measureType
qb:subSlice
skos:Concept
qb:AttributeProperty
qb:observation
sdmx:Concept
qb:MeasureProperty
qb:Observation
sdmx:ConceptRole
qb:CodedProperty
qb:codeList
skos:ConceptScheme
sdmx:FrequencyRole
sdmx:CountRole
sdmx:EntityRole
sdmx:TimeRole
sdmx:MeasureTypeRole
sdmx:NonObsTimeRole
sdmx:IdentityRole
sdmx:PrimaryMeasureRole
sdmx:CodeList

QB4OLAP

- QB4OLAP adds to Data Cube vocabulary (QB) the capability of representing:
  - dimension levels
  - level members
  - rollup relations between levels and level members
  - associating aggregate functions to measures
- It allows to represent OLAP cubes in RDF
- It allows to implement OLAP operators (such as Roll-up, Slice, and Dice) as SPARQL queries directly on this RDF representation

http://publishing-multidimensional-data.googlecode.com/git/index.html

- Unemployment

Mercado laboral

Paro registrado

Paro registrado por CCAA y periodo.

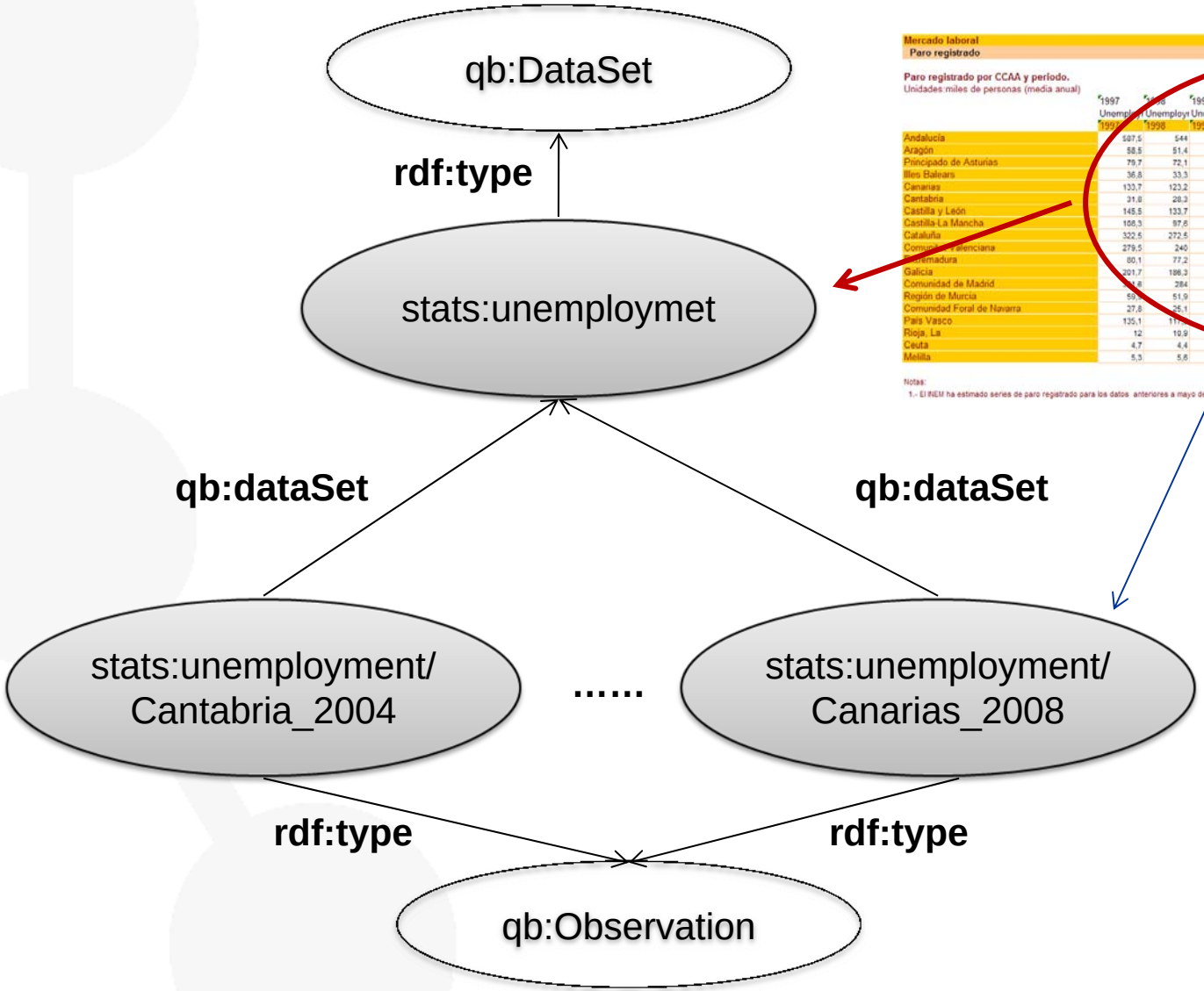
Unidades:miles de personas (media anual)

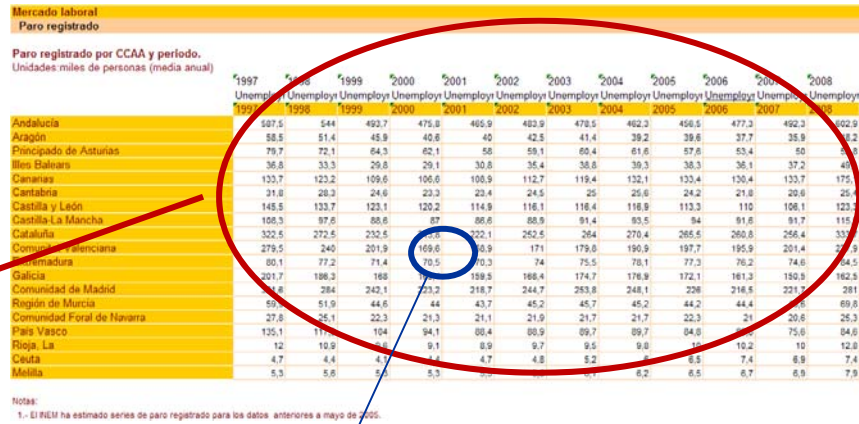
	1997 Unemployr	1998 Unemployr	1999 Unemployr	2000 Unemployr	2001 Unemployr	2002 Unemployr	2003 Unemployr	2004 Unemployr	2005 Unemployr	2006 Unemployr	2007 Unemployr	2008 Unemployr
	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Andalucía	587,5	544	493,7	475,8	465,9	483,9	478,5	462,3	456,5	477,3	492,3	602,9
Aragón	58,5	51,4	45,9	40,6	40	42,5	41,4	39,2	39,6	37,7	35,9	48,2
Principado de Asturias	79,7	72,1	64,3	62,1	58	59,1	60,4	61,6	57,6	53,4	50	52,8
Illes Balears	36,8	33,3	29,8	29,1	30,8	35,4	38,8	39,3	38,3	36,1	37,2	49,1
Canarias	133,7	123,2	109,6	106,6	108,9	112,7	119,4	132,1	133,4	130,4	133,7	175,1
Cantabria	31,8	28,3	24,6	23,3	23,4	24,5	25	25,6	24,2	21,8	20,6	25,4
Castilla y León	145,5	133,7	123,1	120,2	114,9	116,1	116,4	116,9	113,3	110	106,1	123,3
Castilla-La Mancha	108,3	97,6	88,6	87	86,6	88,9	91,4	93,5	94	91,6	91,7	115,5
Cataluña	322,5	272,5	232,5	213,8	222,1	252,5	264	270,4	265,5	260,8	256,4	333,7
Comunitat Valenciana	279,5	240	201,9	169,6	158,9	171	179,8	190,9	197,7	195,9	201,4	277,9
Extremadura	80,1	77,2	71,4	70,5	70,3	74	75,5	78,1	77,3	76,2	74,6	84,5
Galicia	201,7	186,3	168	163,4	159,5	168,4	174,7	176,9	172,1	161,3	150,5	162,5
Comunidad de Madrid	321,6	284	242,1	223,2	218,7	244,7	253,8	248,1	226	216,5	221,7	281
Región de Murcia	59,9	51,9	44,6	44	43,7	45,2	45,7	45,2	44,2	44,4	46,6	69,8
Comunidad Foral de Navarra	27,8	25,1	22,3	21,3	21,1	21,9	21,7	21,7	22,3	21	20,6	25,3
País Vasco	135,1	117,9	104	94,1	88,4	88,9	89,7	89,7	84,8	80,8	75,6	84,6
Rioja, La	12	10,9	9,6	9,1	8,9	9,7	9,5	9,8	10	10,2	10	12,8
Ceuta	4,7	4,4	4,1	4,4	4,7	4,8	5,2	6	6,5	7,4	6,9	7,4
Melilla	5,3	5,6	5,3	5,3	5,5	5,6	6,1	6,2	6,5	6,7	6,9	7,9

Notas:

1.- El INEM ha estimado series de paro registrado para los datos anteriores a mayo de 2005.

Example

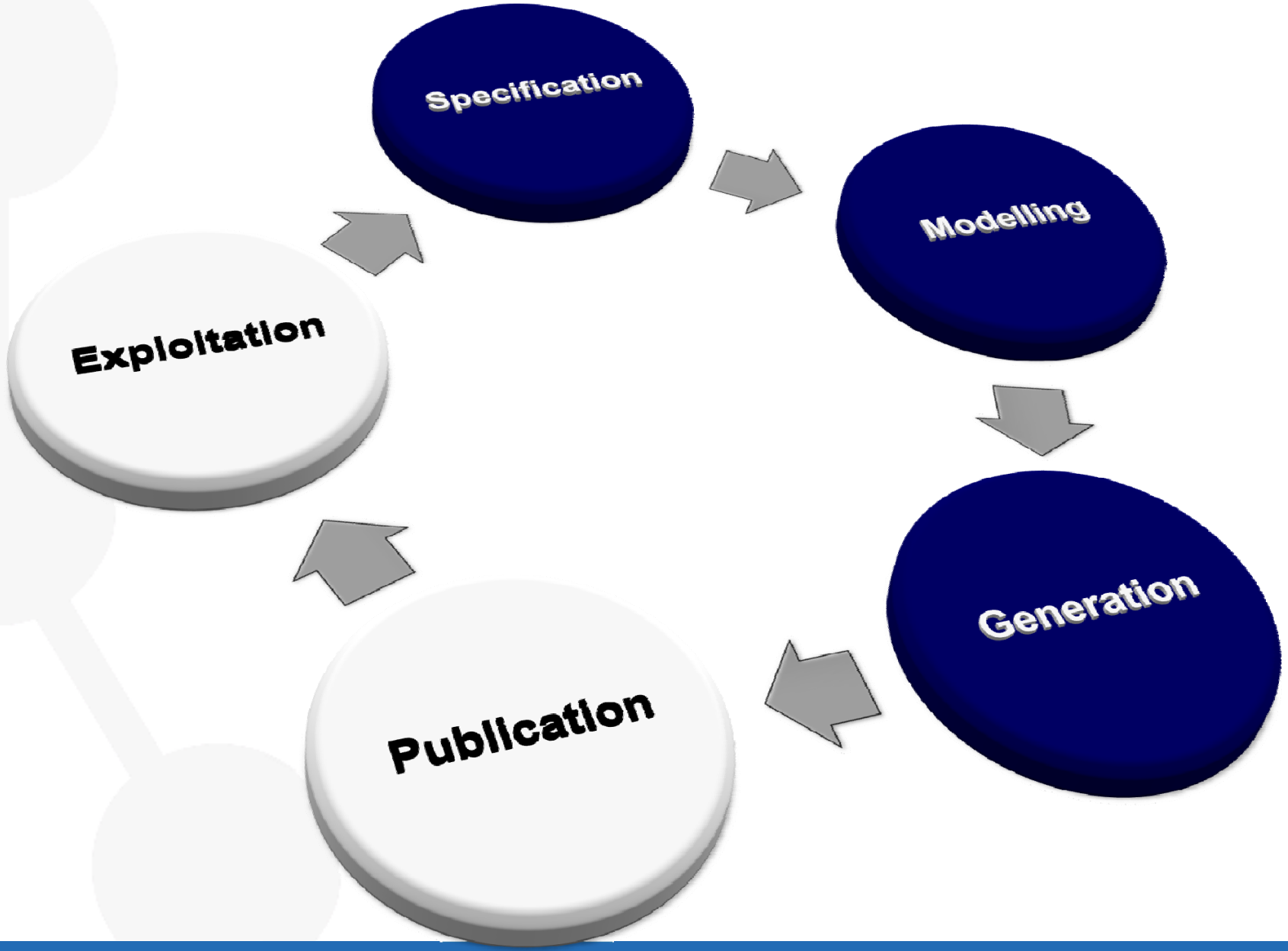



Mercado laboral
Paro registrado

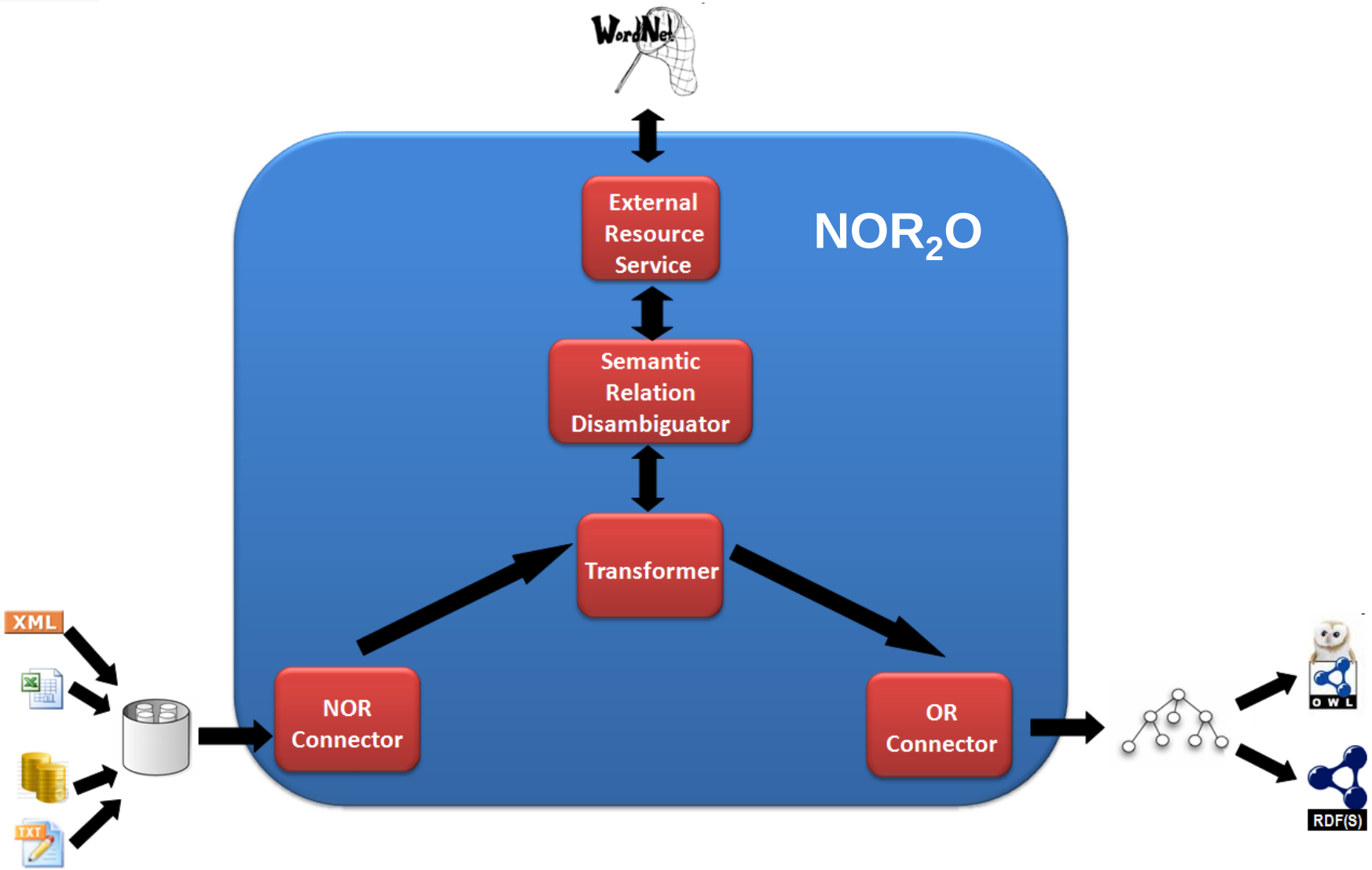
Paro registrado por CCAA y período.
Unidades miles de personas (media anual)

	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Andalucía	597,5	544	493,7	475,8	465,9	483,9	478,5	462,3	456,5	477,3	492,3	602,9
Aragón	58,5	51,4	45,9	40,6	40	42,5	41,4	39,2	39,6	37,7	35,9	48,2
Principado de Asturias	79,7	72,1	64,3	62,1	58	59,1	60,4	61,6	57,6	53,4	50	57
Illes Balears	36,8	33,3	29,8	29,1	30,8	35,4	38,6	39,3	38,3	36,1	37,2	49
Canarias	133,7	123,2	109,6	106,6	106,9	112,7	119,4	132,1	133,4	130,4	133,7	175,1
Cantabria	31,8	28,3	24,6	23,3	23,4	24,5	25	25,6	24,2	21,8	20,6	25,4
Castilla y León	145,5	133,7	123,1	120,2	114,9	116,1	116,4	116,9	113,3	110	106,1	123,3
Castilla-La Mancha	106,3	97,6	88,6	87	86,6	88,9	91,4	93,5	94	91,6	91,7	115
Cataluña	322,5	272,5	232,5	[illegible]	222,1	252,5	264	270,4	266,5	260,8	256,4	337
Comunidad Valenciana	279,5	240	201,9	169,6	170,9	171	179,6	190,9	197,7	195,9	201,4	270,9
Extremadura	80,1	77,2	71,4	70,5	70,3	74	75,5	78,1	77,3	76,2	74,6	84,5
Galicia	201,7	186,3	168	[illegible]	159,5	168,4	174,7	176,9	172,1	161,3	150,5	162,5
Comunidad de Madrid	314,8	284	242,1	223,2	218,7	244,7	253,8	248,1	226	216,5	221,7	281
Región de Murcia	59	51,9	44,6	44	43,7	45,2	45,7	45,2	44,2	44,4	46	69,8
Comunidad Foral de Navarra	27,8	25,1	22,3	21,3	21,1	21,9	21,7	21,7	22,3	21	20,6	25,3
País Vasco	135,1	119	104	94,1	88,4	88,9	89,7	89,7	84,8	79,8	75,6	84,6
Rioja, La	12	10,9	9,6	9,1	8,9	9,7	9,5	9,8	10	10,2	10	12,8
Ceuta	4,7	4,4	4,2	4,4	4,7	4,8	5,2	6	6,5	7,4	6,9	7,4
Melilla	5,3	5,6	5,3	5,3	5,5	5,7	5,7	6,2	6,5	6,7	6,9	7,9

Notas:
1.- El INEM ha estimado series de paro registrado para los datos anteriores a mayo de 2005.

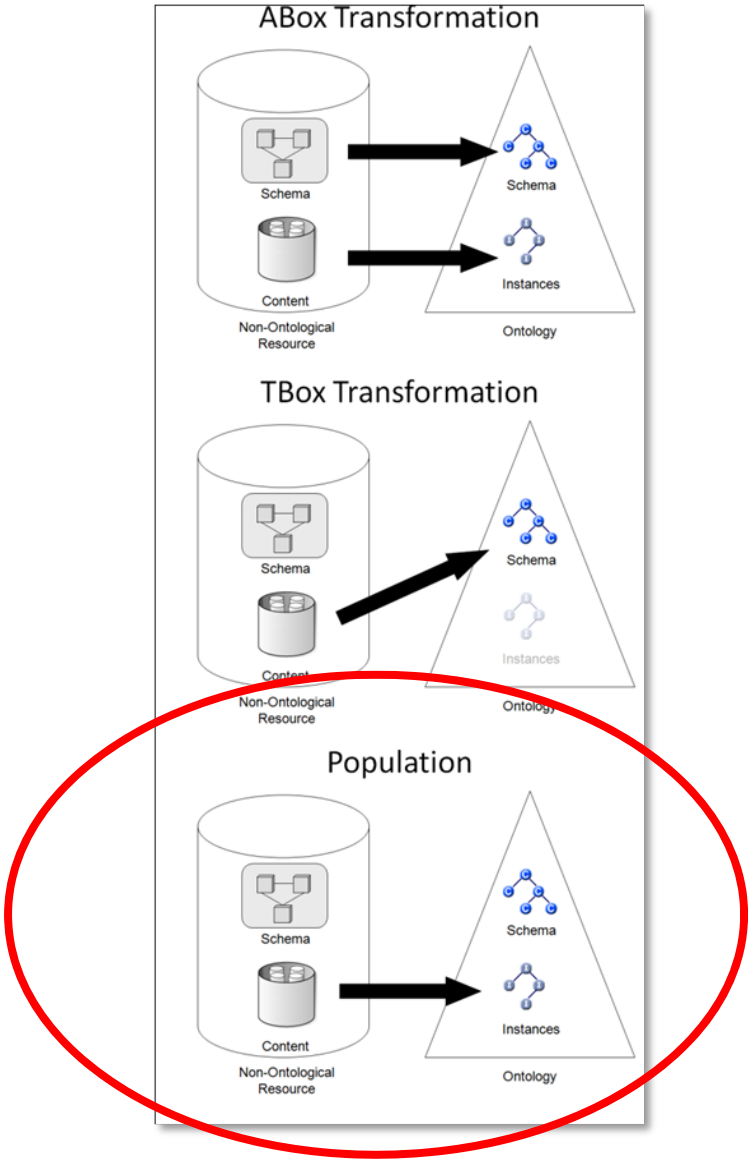

Specification
Modelling
Generation
Publication
Exploitation

NOR$_2$O - Spreadsheets 2 RDF



http://oeg-dev.dia.fi.upm.es/nor2o

Transformation approaches



- 3 configuration files:
  - nor.xml to describe the non ontological resources
  - prnor.xml to describe the transformation
  - or.xml to describe the namespaces of the used ontologies and the name of the resulting RDF

NOR$_2$O – nor.xml

	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Andalucía	587,5	544	493,7	475,8	465,9	483,9	478,5	462,3	456,5	477,3	492,3	602,9
Aragón	58,5	51,4	45,9	43,3	40	42,5	41,4	40,2	39,6	37,7	35,9	49,2
Principado de Asturias	79,7	72,1	67,1	67,1	58	59,1	60,4	61,6	57,1	57,4	50	62,8
Illes Balears	36,8	33,3	29,8	29,1	30,8	35,4	38,8	39,3	38,3	38	37,2	49,1
Canarias	133,7	123,2	109,6	106,6	108,9	112,7	119,4	132,1	133,4	130,4	123,7	175,1
Cantabria	31,8	28,3	24,6	23,3	23,4	24,5	25	25,6	24,2	21,8	20,6	25,4
Castilla y León	145,5	133,7	123,1	120,2	114,9	116,1	116,4	116,9	113,3	110	106,1	123,3
Castilla-La Mancha	108,3	97,6	88,6	87	86,6	88,9	91,4	93,5	94	91,6	91,7	115,5
Cataluña	322,5	272,5	232,5	213,8	222,1	252,5	264	270,4	265,5	260,8	256,4	333,1
Comunitat Valenciana	279,5	240	201,9	169,6	168,9	171	179,8	190,9	197,7	195,9	201,4	277,1
Extremadura	80,1	77,2	71,4	70,5	70,3	74	75,5	78,1	77,3	76,2	74,6	84,1
Galicia	201,7	186,3	168	163,4	159,5	168,4	174,7	176,9	172,1	161,3	150,5	162,8
Comunidad de Madrid	321,6	284	242,1	223,2	218,7	244,7	253,8	248,1	226	216,5	221,7	291
Región de Murcia	59,9	51,9	44,6	44	43,7	45,2	45,7	45,2	44,2	44,4	46,6	69,8
Comunidad Foral de Navarra	27,8	25,1	22,3	21,3	21,1	21,9	21,7	21,7	22,3	21	20,5	25,3
País Vasco	135	117,9	104	94,1	88,4	88,9	89,7	89,7	84,8	80,8	78,5	84,6
Rioja, La	12	10,9	9,6	9,1	8,9	9,7	9,6	9,8	10	10,2	10	12,8
Ceuta	4,7	4,4	4,1	4,4	4,7	4,8	5,2	6	6,5	7,4	6,9	7,4
Melilla	5,3	5,6	5,3	5,6	5,5	5,6	5,1	5,1	6,5	6,7	6,9	7,9

```
Spanish Employment">
    <SchemaEntities>
        <SchemaEntity name="Location">
            <Attribute name="Name" valueFrom="pcaxis.[A9:A25]" type="string"/>
        </SchemaEntity>
        <SchemaEntity name="Dataset">
            <Attribute name="Name" valueFrom="pcaxis.[B7:M7]" type="string"/>
        </SchemaEntity>
        <SchemaEntity name="Year">
            <Attribute name="Name" valueFrom="pcaxis.[B6:M6]" type="string"/>
        </SchemaEntity>
        <SchemaEntity name="UnemploymentByLocationInPeriod" type="Nary">
            <Attribute name="hasValue" valueFrom="pcaxis.[B9:M25]" type="string"/>
            <Relation name="hasLocation" usingSpreadSheetColumn="A" destination="Location"/>
            <Relation name="inPeriod" usingSpreadSheetRow="6" destination="Year"/>
            <Relation name="dataset" usingSpreadSheetRow="7" destination="Dataset"/>
        </SchemaEntity>
    </SchemaEntities>
</Schema>
<DataModel>
    <GenericDataModel/>
</DataModel>
<Implementation>
    <Spreadsheet type="ms" file="Paro.xls"/>
</Implementation>
</NOR>
```

NOR$_2$O – prnor.xml

- How to transform the resource

```
<Prnor identifier="PR-NOR-CLLD-01" transformationApproach="Population">
    <Class from="UnemploymentByLocationInPeriod" identifier="cube:Observation">
        <ObjectProperty from="hasLocation" to="property:geoArea"/>
        <ObjectProperty from="inPeriod" to="dimension:refPeriod"/>
        <ObjectProperty from="dataset" to="cube:dataSet"/>
        <DataTypeProperty from="[hasValue]" to="property:Unemployment" type="http://www.w3.org/2001/XMLSchema#double"/>
        <Individual from="UnemploymentByLocationInPeriod" identifier="unemployment/.[hasLocation]._.[inPeriod]"/>
    </Class>
<Class from="Dataset" identifier="cube:DataSet">
        <DataTypeProperty from="[Name]" to="nombre" type="http://www.w3.org/2001/XMLSchema#string"/>
        <Individual from="Dataset" identifier="[Name]"/>
</Class>
<Class from="Year" identifier="Year">
        <DataTypeProperty from="[Name]" to="nombre" type="http://www.w3.org/2001/XMLSchema#string"/>
        <Individual from="Year" identifier="year/.[Name]"/>
</Class>
<Class from="Location" identifier="Region">
        <DataTypeProperty from="[Name]" to="nombre" type="http://www.w3.org/2001/XMLSchema#string"/>
        <Individual from="Location" identifier="geoes:ComunidadAut%C3%B3noma/.[Name]"/>
</Class>
</Prnor>
```

Reusing GeoLinkedData URIs!

NOR$_2$O – or.xml

- To describe the namespaces and the resulting RDF

```
<Or name="Unemployment Index" ontologyURI="http://stats.ull.es/resource/"
  ontologyFile="paro.rdf" implementation="OWL" alreadyExist="no" separator="">
<prefix name="geoos" uri="http://geo.linkeddata.es/resource/" />
<prefix name="cube" uri="http://purl.org/linked-data/cube#" />
<prefix name="property" uri="http://stats.ull.es/property/" />
<prefix name="dimension" uri="http://purl.org/linked-data/sdmx/2009/dimension#" />
<prefix name="rdf" uri="http://www.w3.org/1999/02/22-rdf-syntax-ns#" />
<prefix name="year" uri="http://reference.data.gov.uk/id/year/" />
</Or>
```

NOR_2O – Exercise

- **What you need:**
  - Java 1.5 o superior
- **What to do:**
  - Download NOR2O

    http://oeg-dev.dia.fi.upm.es/nor2o/#download
  - Extract the files in your workspace folder.
  - The current distribution comes with three examples in folders **example1**, **example2** and **example3**.
  - Files in the wiki
    - “Paro” case: Unemployement for each city by date, genre and age
    - “Atlas” case: Number of health centers for each city by year

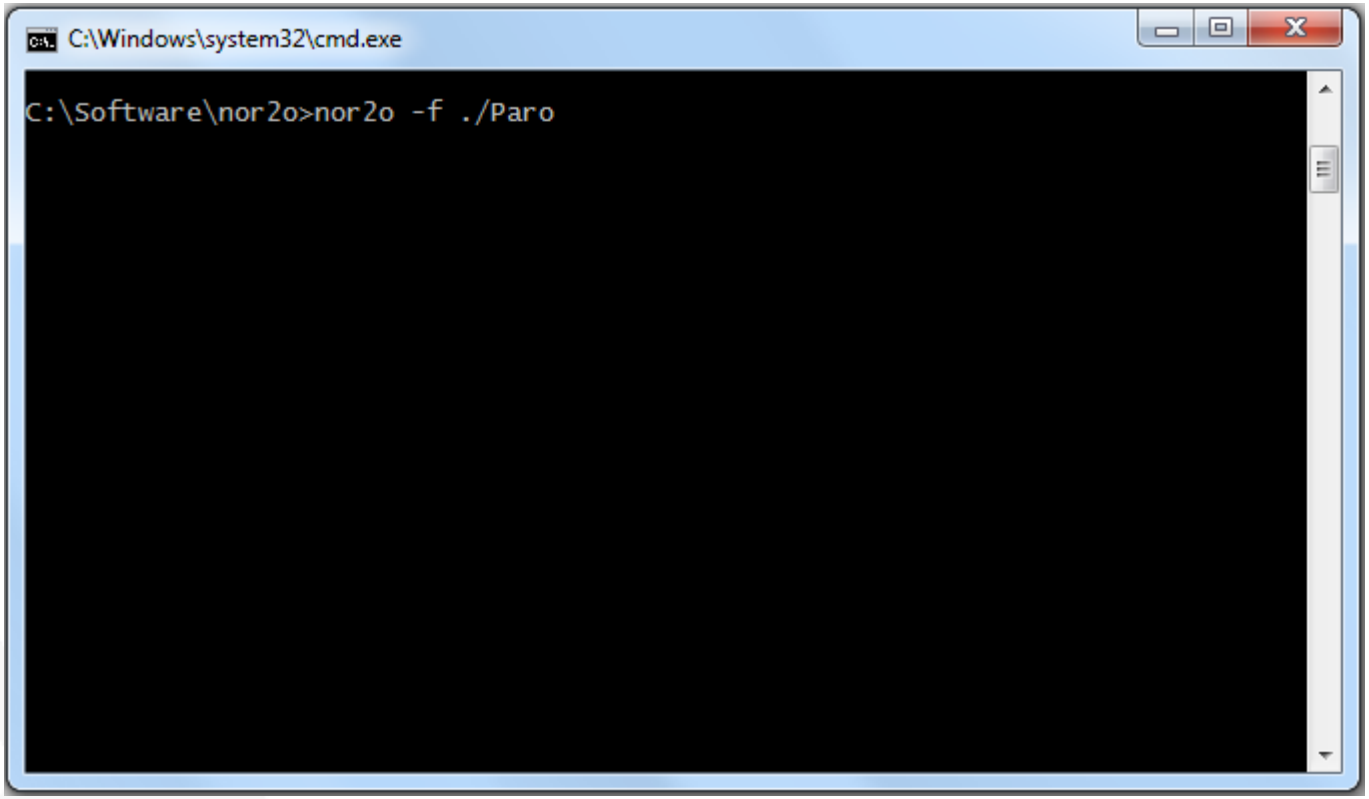
NOR$_2$O – Exercise

- **What you need:**
  - Java 1.5 o superior
- **What to do:**
  - Download NOR2O

    http://oeg-dev.dia.fi.upm.es/nor2o/#download
  - Extract the files in your workspace folder.
  - The current distribution comes with three examples in folders example1, example2 and example3.
  - Files in the wiki
    - “Paro” case: Unemployement for each city by date, genre and age
    - “Atlas” case: Number of health centers for each city by year
  - Run nor2o from the command line
  - Wait until it finishes and check the generated RDF file

NOR$_2$O – Exercise

- Run nor2o from the command line



Data cubes and NOR_2O

Esther Lozano, Boris Villazón-Terrazas, Oscar Corcho
Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo sn, 28660 Boadilla del Monte, Madrid
http://www.oeg-upm.net
elozano@fi.upm.es
Phone: 34.91.3366605, Fax: 34.91.3524819