


ER 2015

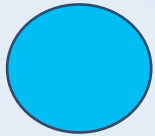
Carlo Batini
University of Milano
Bicocca
batini@disco.unimib.it

Tutorial on

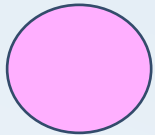
**COMSODE -
A Methodology
for publishing
datasets as open
data**

- 
- *The black ribbon represents the mourning of the COMSODE consortium caused by loss of Mr Ivan Hanzlik, the Exploitation Manager of COMSODE, who succumbed on Saturday 2nd October 2015 to a serious illness he bravely faced for last year and a half.*
 - *He continued to work on the project till last days of his life, being a great source of inspiration for all of us.*

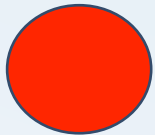
Lot of material for 1 + 1/2 hour...
Three types of slides



Detailed



Fast



Very Fast

Contents



1. Introduction to Open Data and the COMSODE Publication Platform
2. Usage of COMSODE Techniques
3. The COMSODE Methodology: generalities
4. Methodology for one dataset
5. Methodology for multiple datasets
6. Social value of open data

Contents



1. **Introduction to Open Data and the COMSODE Publication Platform**
2. Usage of COMSODE Techniques
3. The COMSODE Methodology: generalities
4. Methodology for one dataset
5. Methodology for multiple datasets
6. Social value of open data

Open Data: a definition

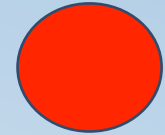


According to the Open Knowledge Foundation (2012) **open data** “is

data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike.

There are two main dimensions of **openness** of open data: **legal** and **technological**.

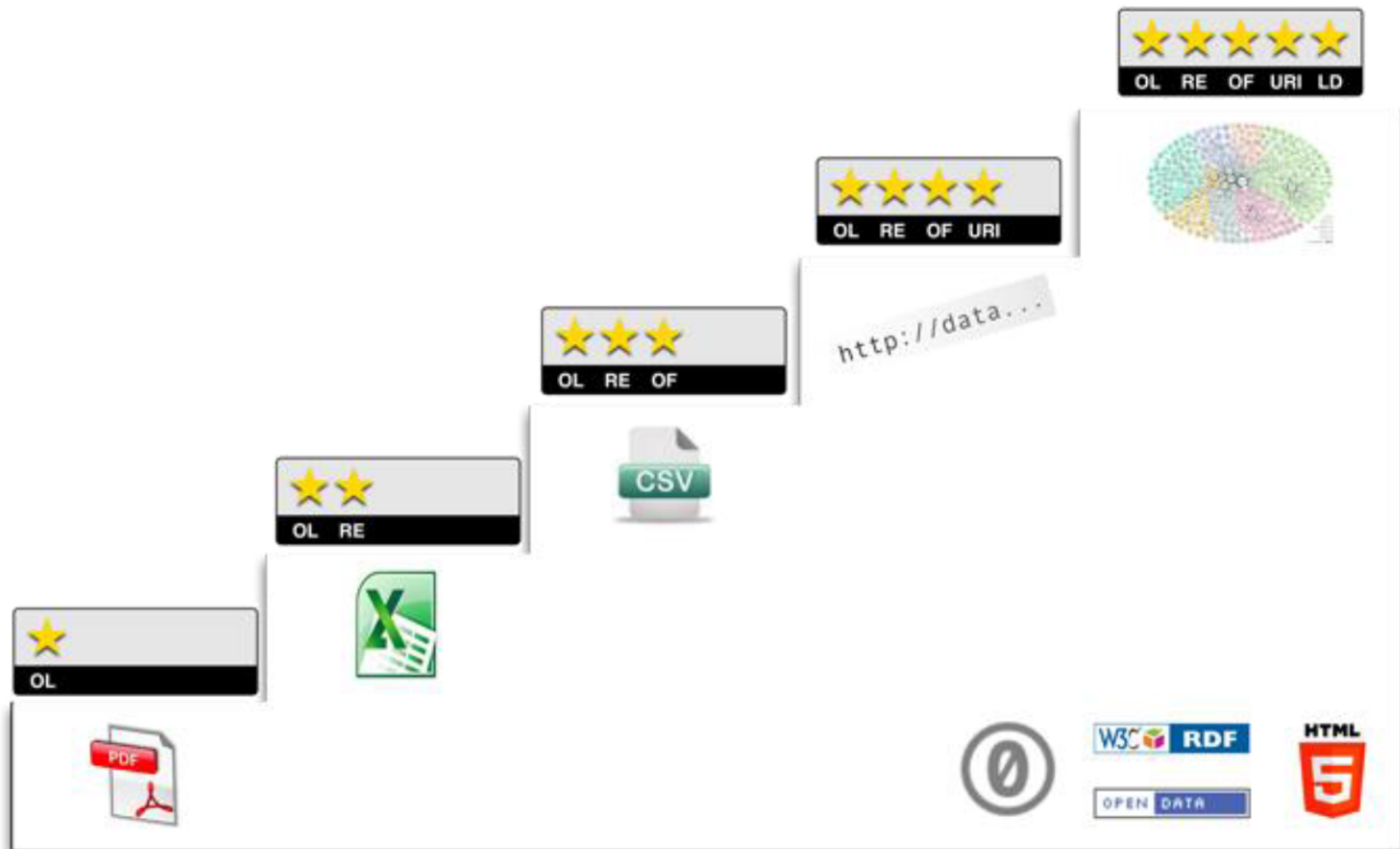
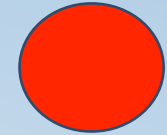
Open Data: properties



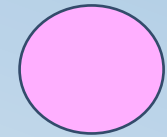
open government data should be:

1. complete,
2. primary,
3. timely,
4. easily accessible,
5. machine readable,
6. non-discriminating,
7. using commonly owned (open) standards,
8. legally open,
9. permanent and
10. non-limiting reuse by fees.

Open Data: the five *

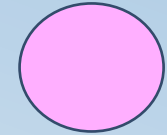


Open Data: the five *



Openness level	Short description
1*	<ul style="list-style-type: none">• Data is available under an open licence.• Any format is used.
2*	<ul style="list-style-type: none">• Data is available under an open licence.• Structured data format is used, e.g. MS Excel.
3*	<ul style="list-style-type: none">• Data is available under an open licence.• Structured and open data format is used, e.g. CSV.
4*	<ul style="list-style-type: none">• Data is available under an open licence.• Structured and open data format is used.• Objects are identified with URIs.
5*	<ul style="list-style-type: none">• Data is available under an open licence.• Structured and open data format is used.• Objects are identified with URIs.• Data is linked to other data in order to provide context.

Open Data: the five *



Openness level	Short description
1*	<ul style="list-style-type: none">• Data is available under an open licence.• Any format is used.
2*	<ul style="list-style-type: none">• Data is available under an open licence.• Structured data format is used, e.g. MS Excel.
3*	<ul style="list-style-type: none">• Data is available under an open licence.• Structured and open data format is used, e.g. CSV.
4*	<ul style="list-style-type: none">• Data is available under an open licence.• Structured and open data format is used.• Objects are identified with URIs.
5*	<ul style="list-style-type: none">• Data is available under an open licence.• Structured and open data format is used.• Objects are identified with URIs.• Data is linked to other data in order to provide context.

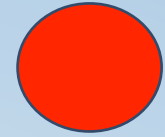
Main Components of COMSODE - o



1. Open Data Node (ODN) **Publication Platform**
2. Unified Views & related **Techniques**
3. Open data Publication **Methodology**
4. **Search** by Strategy **Platform**
5. Open Data **Services**

Main Components of COMSODE - 1

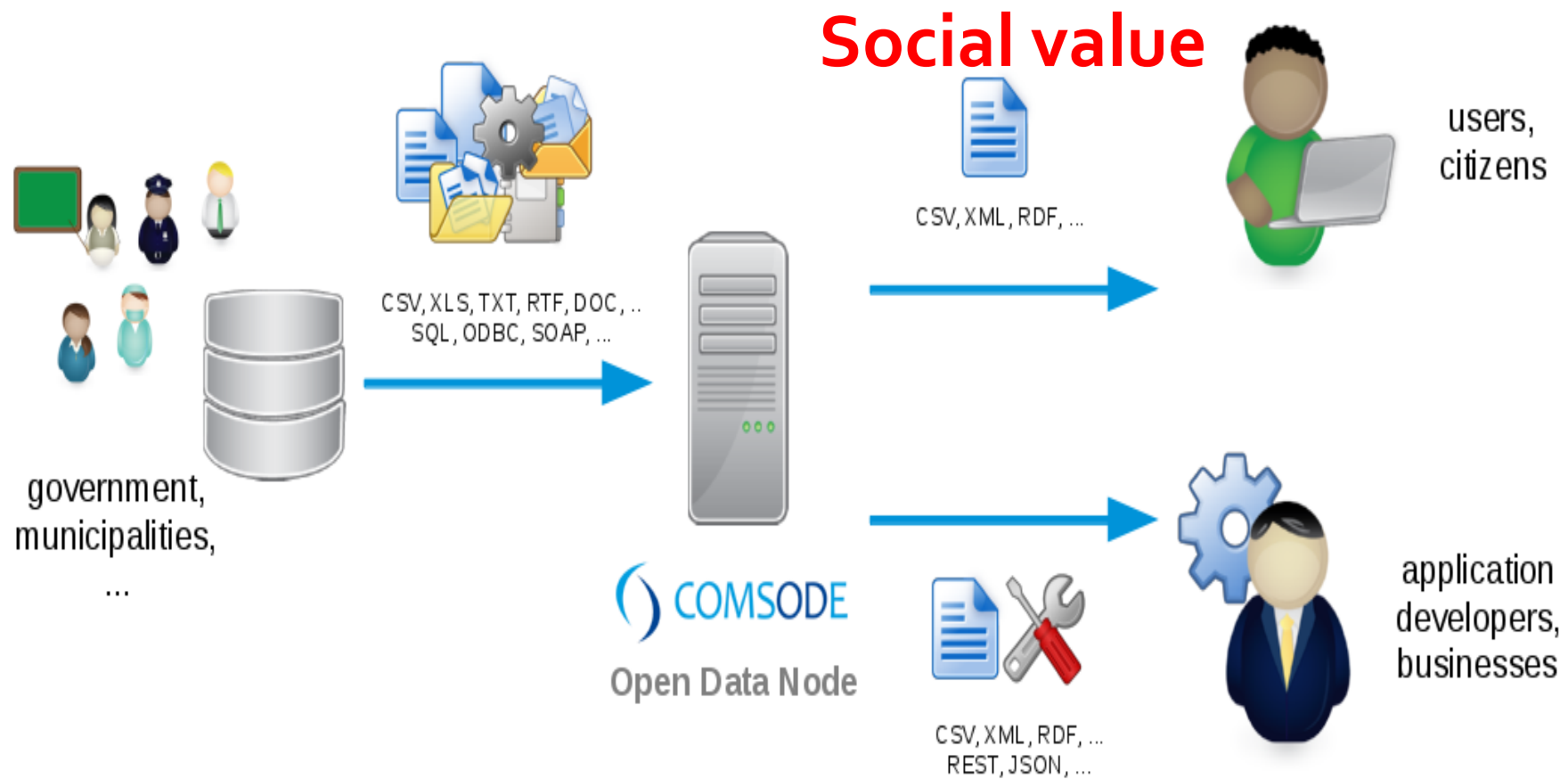
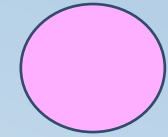
Open Data Node (ODN) publication platform



ODN is **open source software platform** that provides user with tools and functions for effective, sustainable long-term publishing of open data. It **supports entire process of open data publication**, management and exchange.

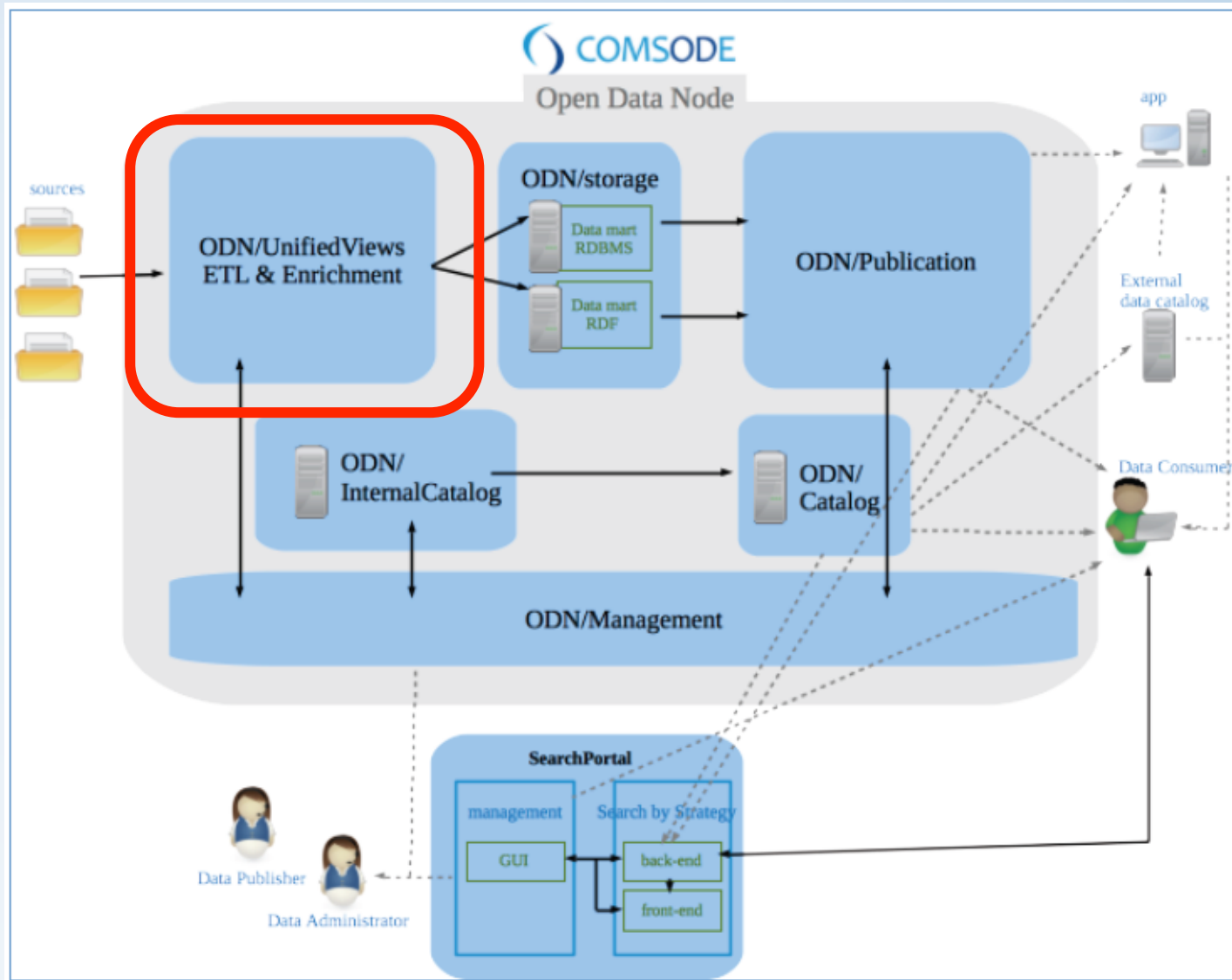
ODN **is equipped with a set of APIs** allowing 3rd parties (SMEs) to build applications on top of ODN.

The Open Data Life Cycle



Economic value

Architecture of Open Data Node Publication Platform



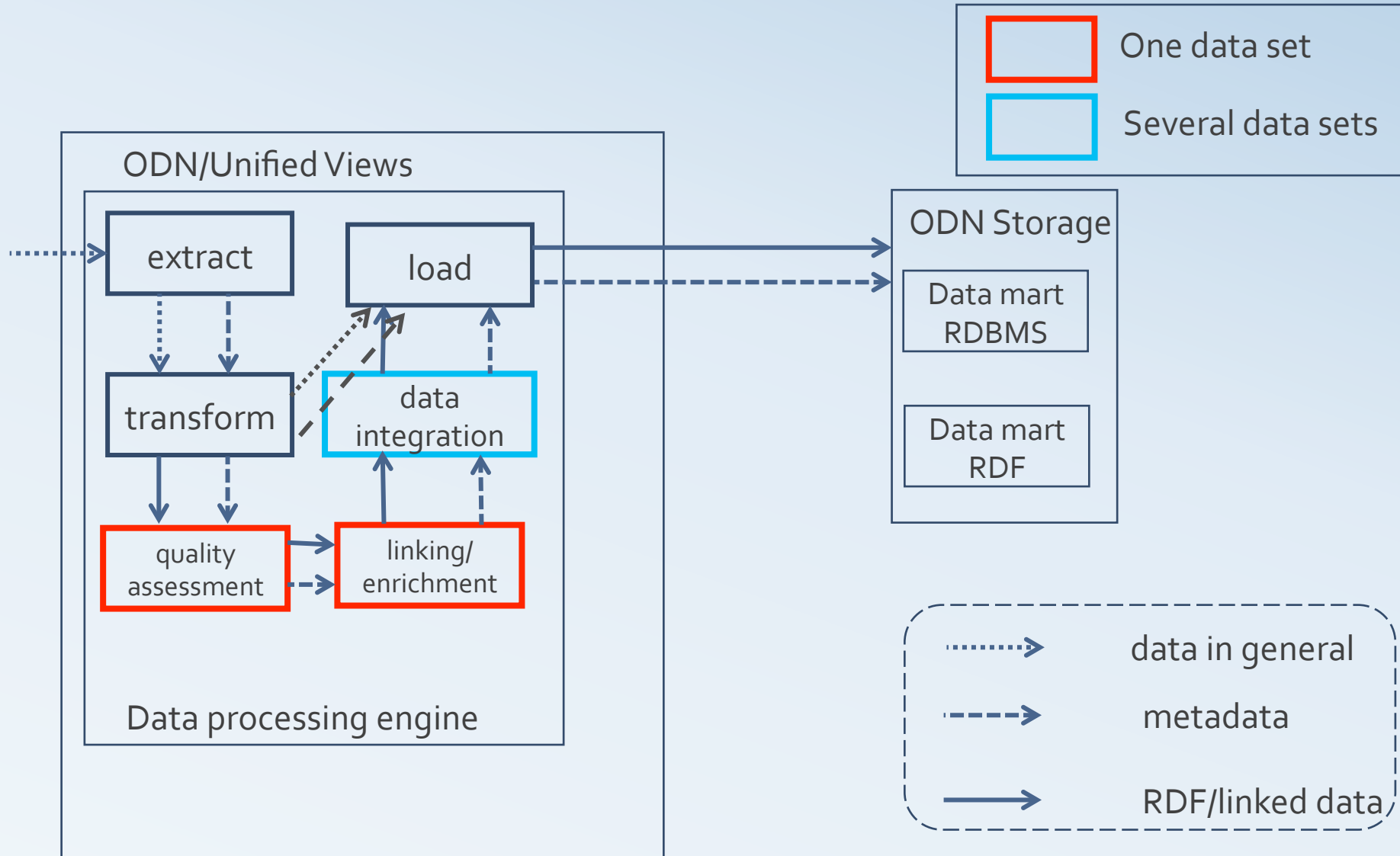
Main Components of COMSODE-2

Unified Views

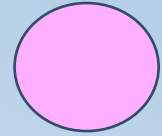
UV is an **Extract-Transform-Load** (ETL) **framework** designed for sustainable data processing.

UV was originally developed as a student project at Charles University in Prague and now it is maintained by EEA (Slovakia, Czech Republic), Semantic Web Company (Austria), Semantica (Czech Republic) and Department of software engineering of Charles University Prague

More on ETL & Enrichment



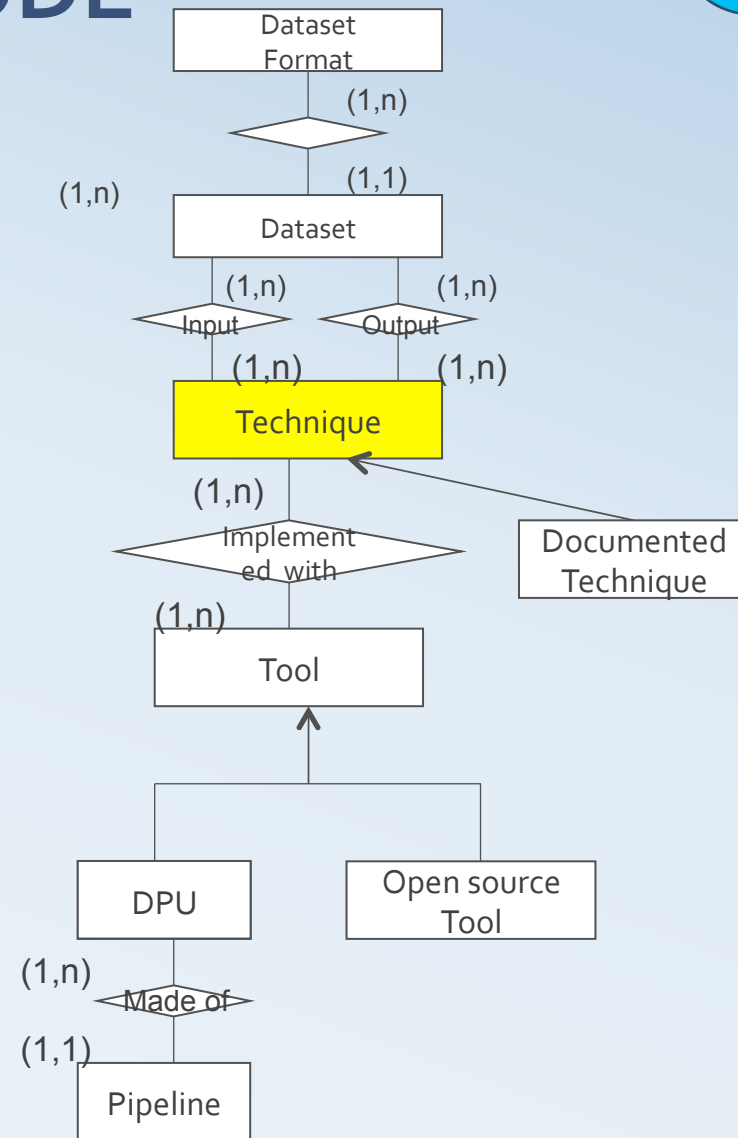
ACRONYMS in the following



extract	EXTR
transform	TRANS
load	LOAD
quality assessment	QUAL
linking/ enrichment	ENR
Data& Schema integration	INT

Techniques in COMSODE

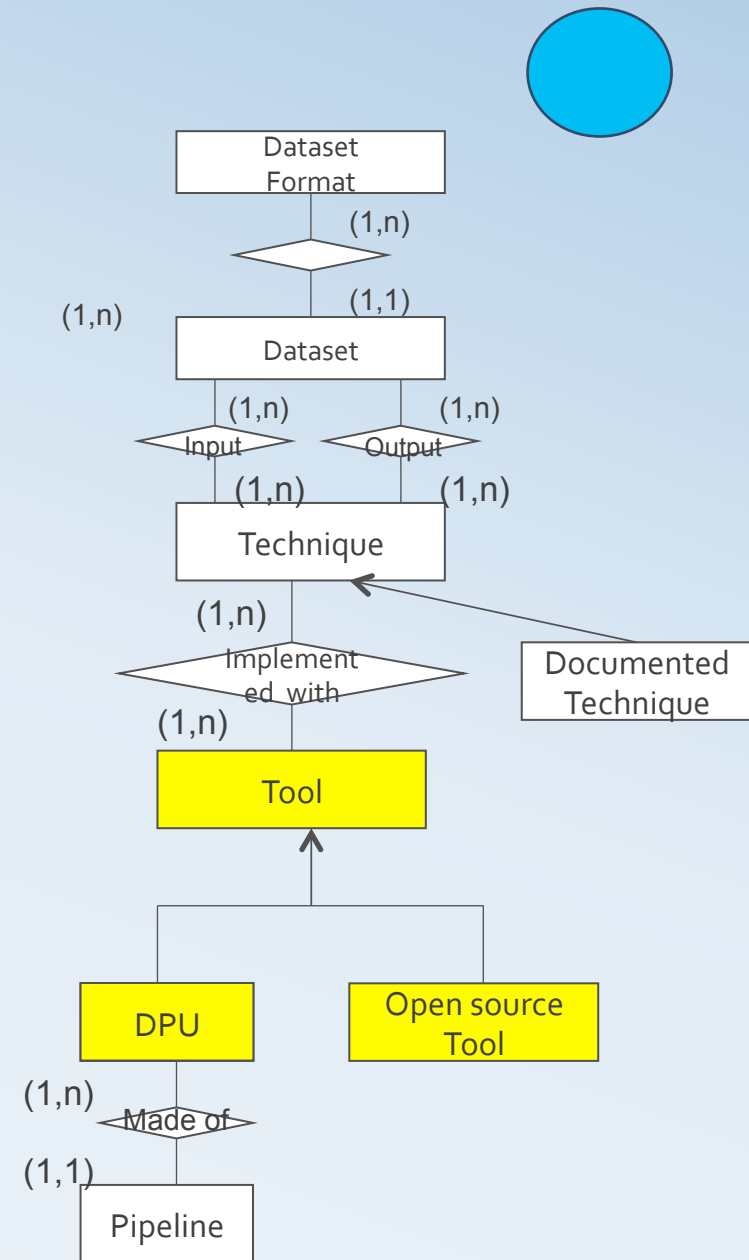
Technique – any algorithm, manual task, software application that can help in the COMSODE life cycle



Techniques in COMSODE

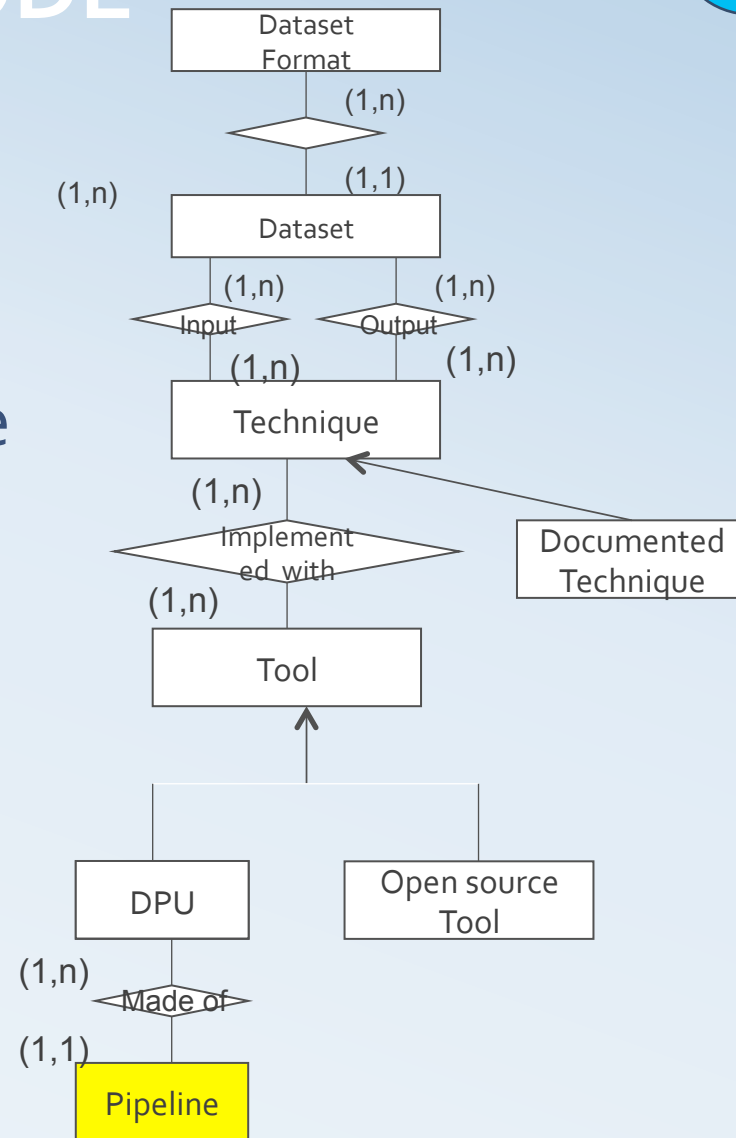
Tool – Technique for which a software application exists.
Tools can be

1. **Data Processing Units (DPUs)** implemented and in a stable version (these are **strongly coupled** tools in the ODN)
2. **Open Source Tools** available as **open source** applications (these are **loosely coupled** tools)



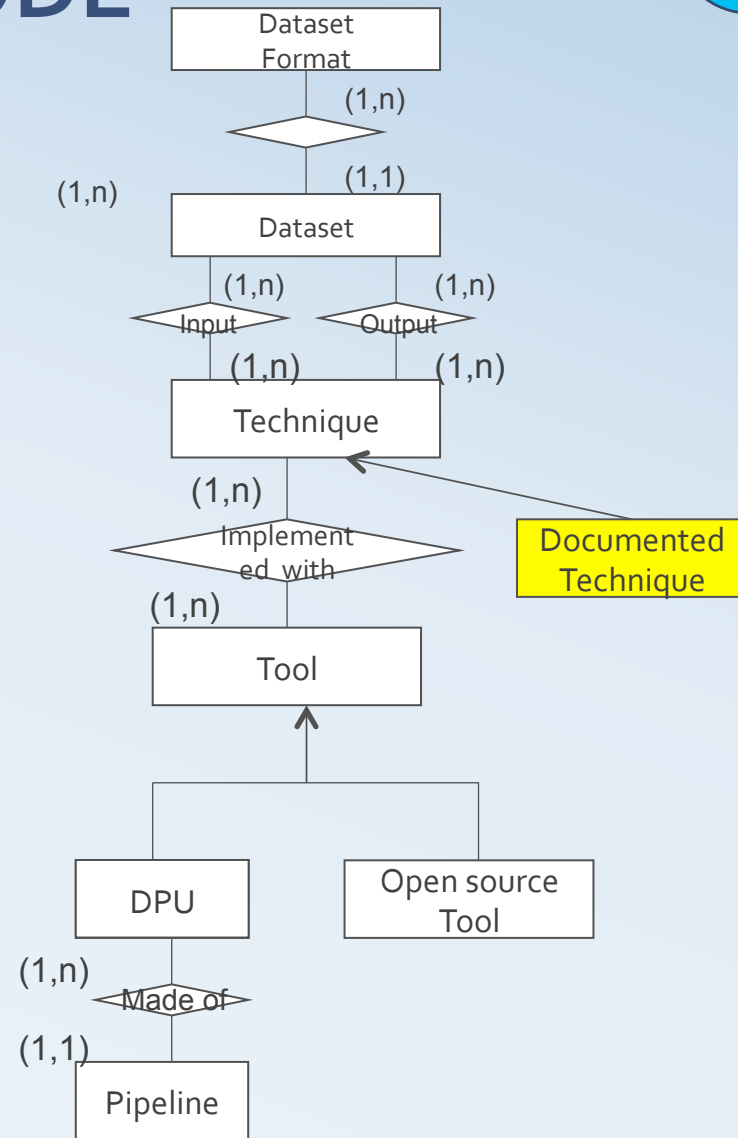
Techniques in COMSODE

Pipeline – workflow made of DPUs that addresses a set of tasks of the COMSODE methodology





Documented technique –



Main Components of COMSODE - 3

Open Data publication methodology

1. For publishing datasets as open data

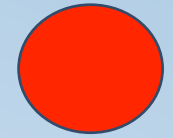
Methodology for publishers (mainly public bodies) about the phases and steps needed for publishing Open Data. It starts from the beginning of the publication activity (“what and why I must publish as open data?”) to the result (“we have dataset suitable for publishing”).

2. For deployment and usage of ODN tools and data

Methodology for publishers (mainly public bodies) wanting to publish their own data using COMSODE publication platform (ODN), techniques and data.

Main Components of COMSODE - 4

Search by Strategy platform

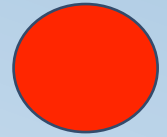


A platform for **easy creation of highly specialized searches** on published data (datasets).

It provide users with more effective ways of exploring data thanks to search capabilities designed on purpose for the use-case at hand, without requiring high technical skills.

Main Components of COMSODE - 5

Open Data Services



Services based on knowledge and experiences gained on project:

- Consulting – methodology, standards, strategy, plan, architecture
- SW development and integration – based on ODN, UnifiedViews and Search by Strategy solution
- Implementation services - installation, configuration, integration, sizing, security issues
- Support, Administration, Hosting of the solution
- Trainings

The main focus of the Tutorial



1. Open Data Node (ODN) **Publication Platform**
2. Unified Views & related **Techniques**
3. Open data Publication **Methodology**
4. **Search** by Strategy **Platform**
5. Open Data **Services**

Contents



1. Introduction to Open Data and the COMSODE Publication Platform
2. **Usage of COMSODE Techniques**
3. The COMSODE Methodology: generalities
4. Methodology for one dataset
5. Methodology for multiple datasets
6. Social value of open data



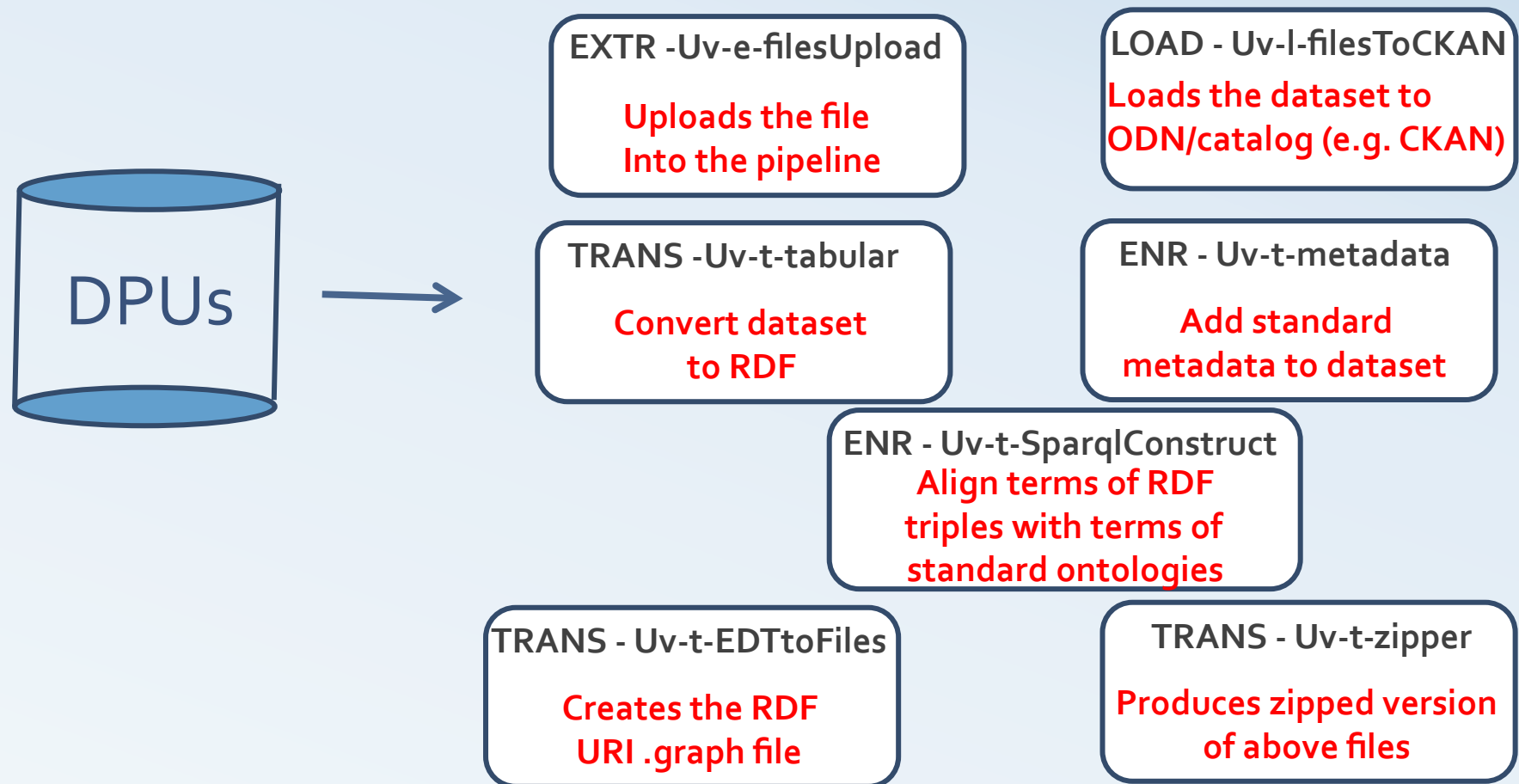
Unified Views **Data Processing Units**

An example of **pipeline** design

An example of pipeline design

Goal: Transformation of a CSV data set into RDF and alignment with a standard ontology

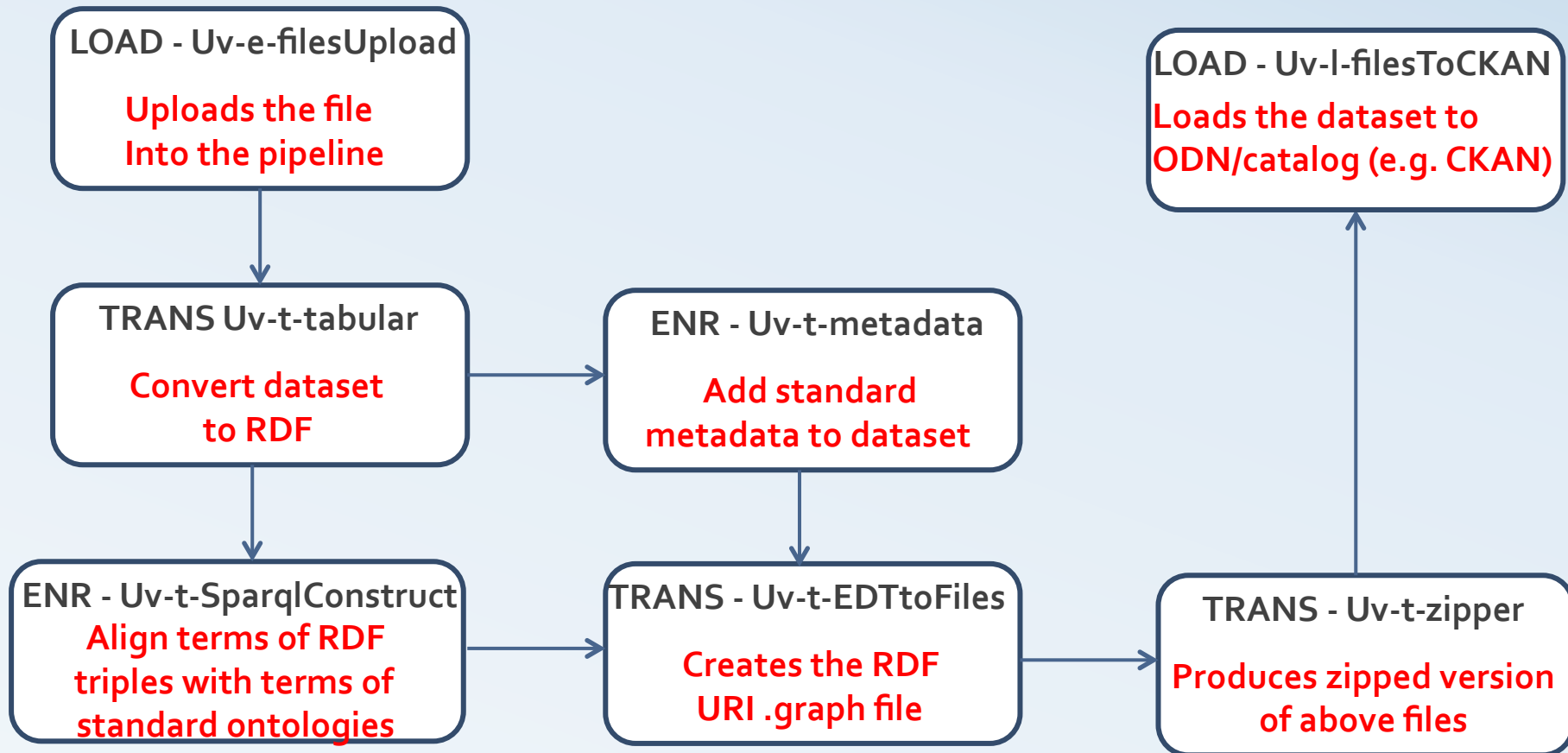
1. Choice of DPU's



An example of pipeline design

Goal: Transformation of a CSV data set into RDF and alignment with a standard ontology

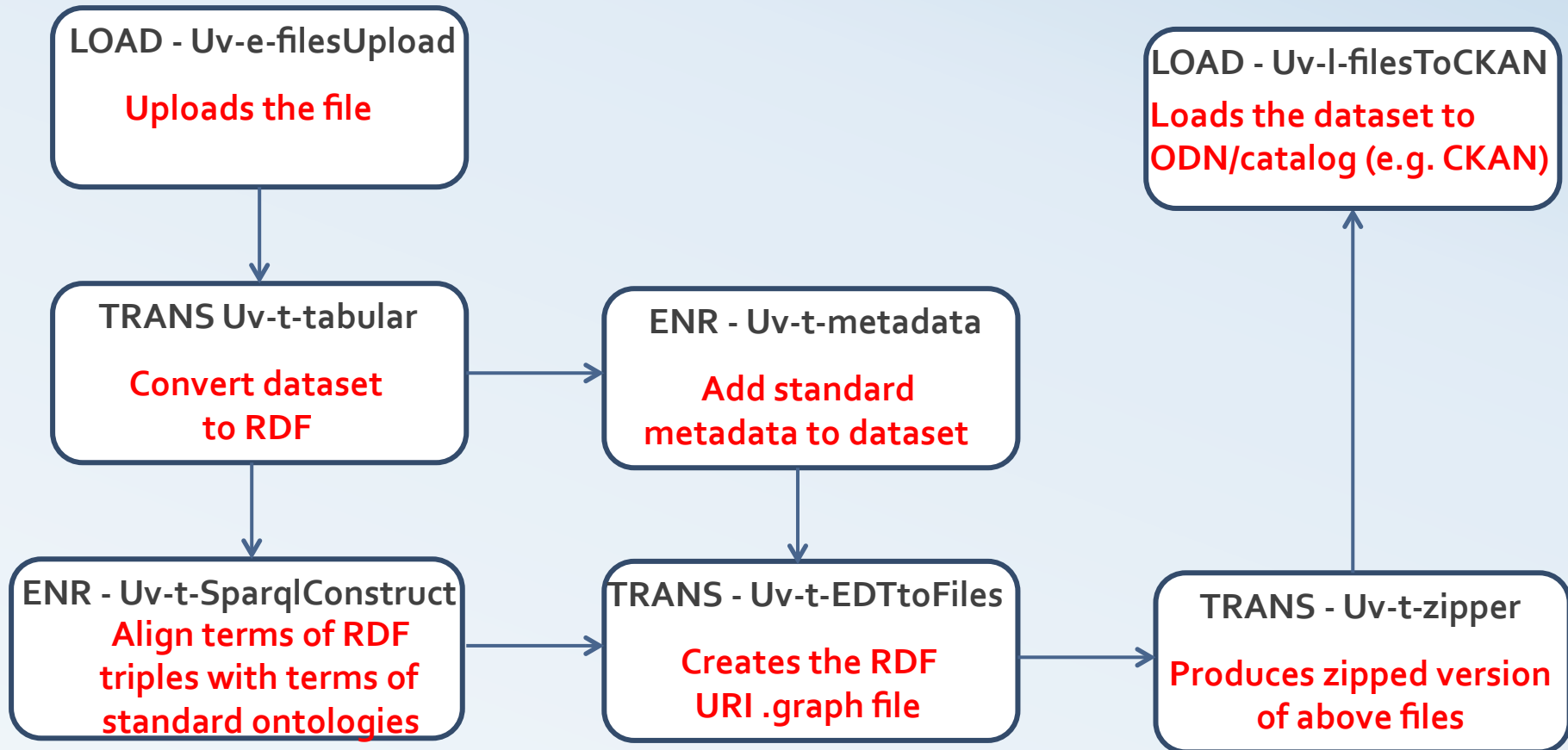
2. Workflow design



An example of pipeline design

Goal: Transformation of a CSV data set into RDF and alignment with a standard ontology

2. Workflow design



Contents

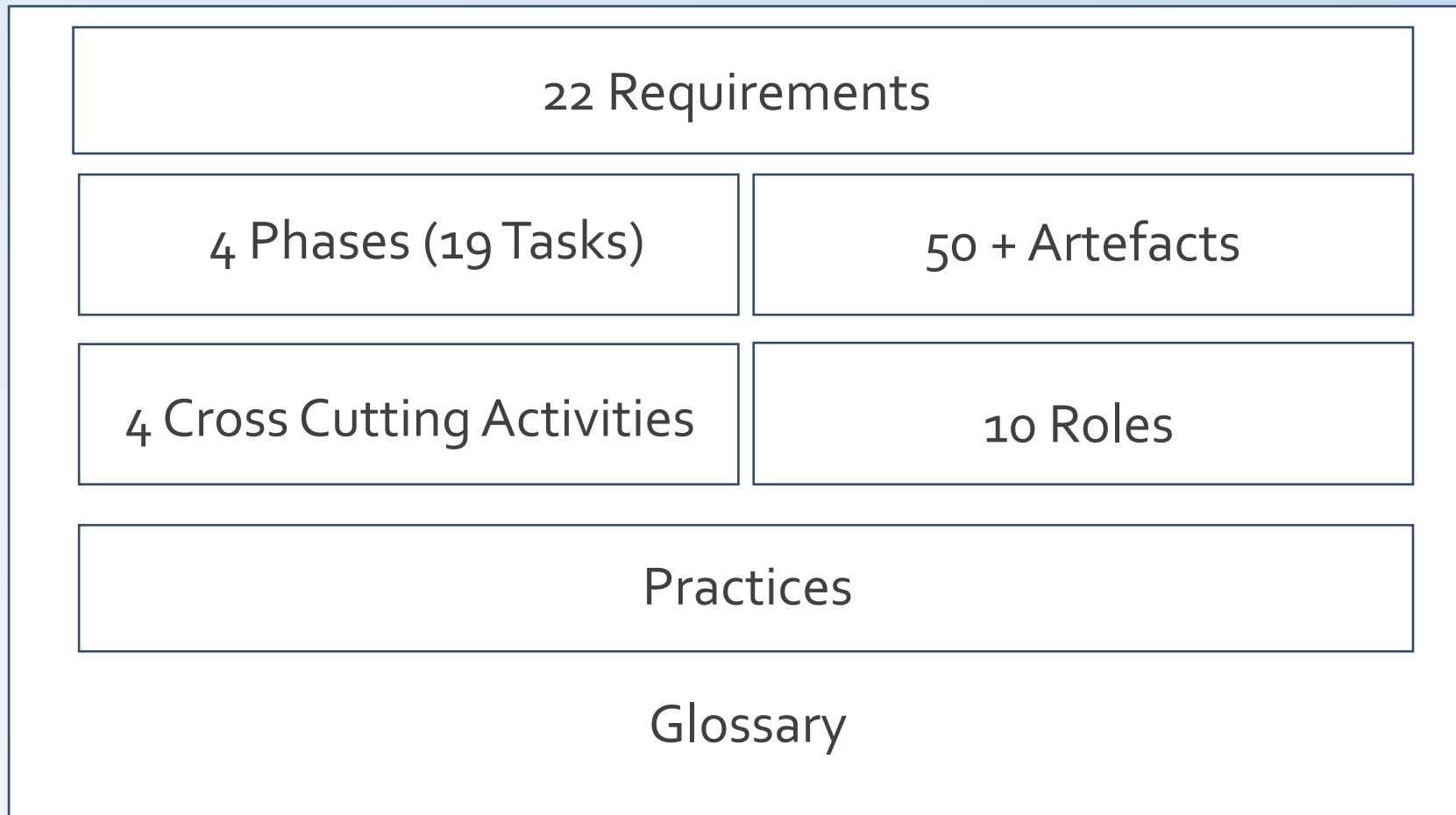


1. Introduction to Open Data and the COMSODE Publication Platform
2. Usage of COMSODE Techniques
3. **The COMSODE Methodology: generalities**
4. Methodology for one dataset
5. Methodology for multiple datasets
6. Social value of open data

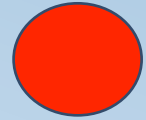


The **COMSODE** Methodology and related DPU's & Techniques

The COMSODE Methodology



Requirements



RQ1- Definition of roles

RQ2 - Assessment of demand for OGD

RQ3 - Selection and prioritization of datasets

RQ4 - OGD benefits assessment

RQ5 - Effort and costs estimation

RQ6 - Recommendation about fees

RQ7 - Ensuring compliance with the legislation

RQ8 - Risk analysis

RQ9 - Licensing

RQ10 - Reuse of already published datasets

RQ11 - Recommended data formats

RQ12 - Interlinking of related datasets

RQ13 - ICT impact assessment

RQ14 - OGD publication process

RQ15 – Data cataloguing

RQ16 – Data quality assurance

RQ17 - Ensuring easy access to datasets

RQ19 - Communication strategy

RQ20 - Independence on the central data portal

RQ21 - Recommended software

RQ22 - Public sector bodies of different size should be taken into consideration



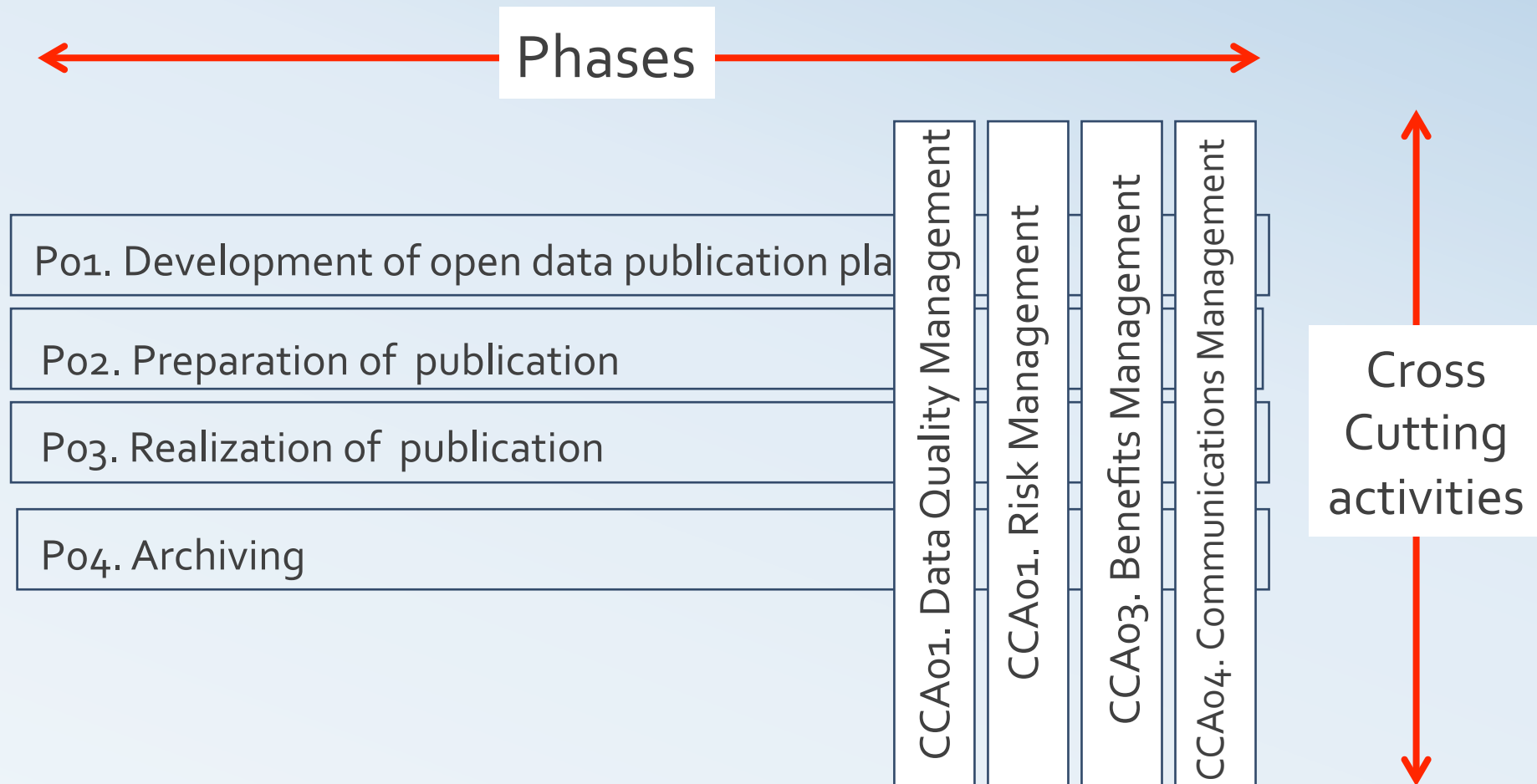
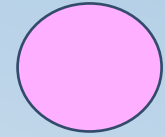
Roles in the Methodology

Roles in the Methodology

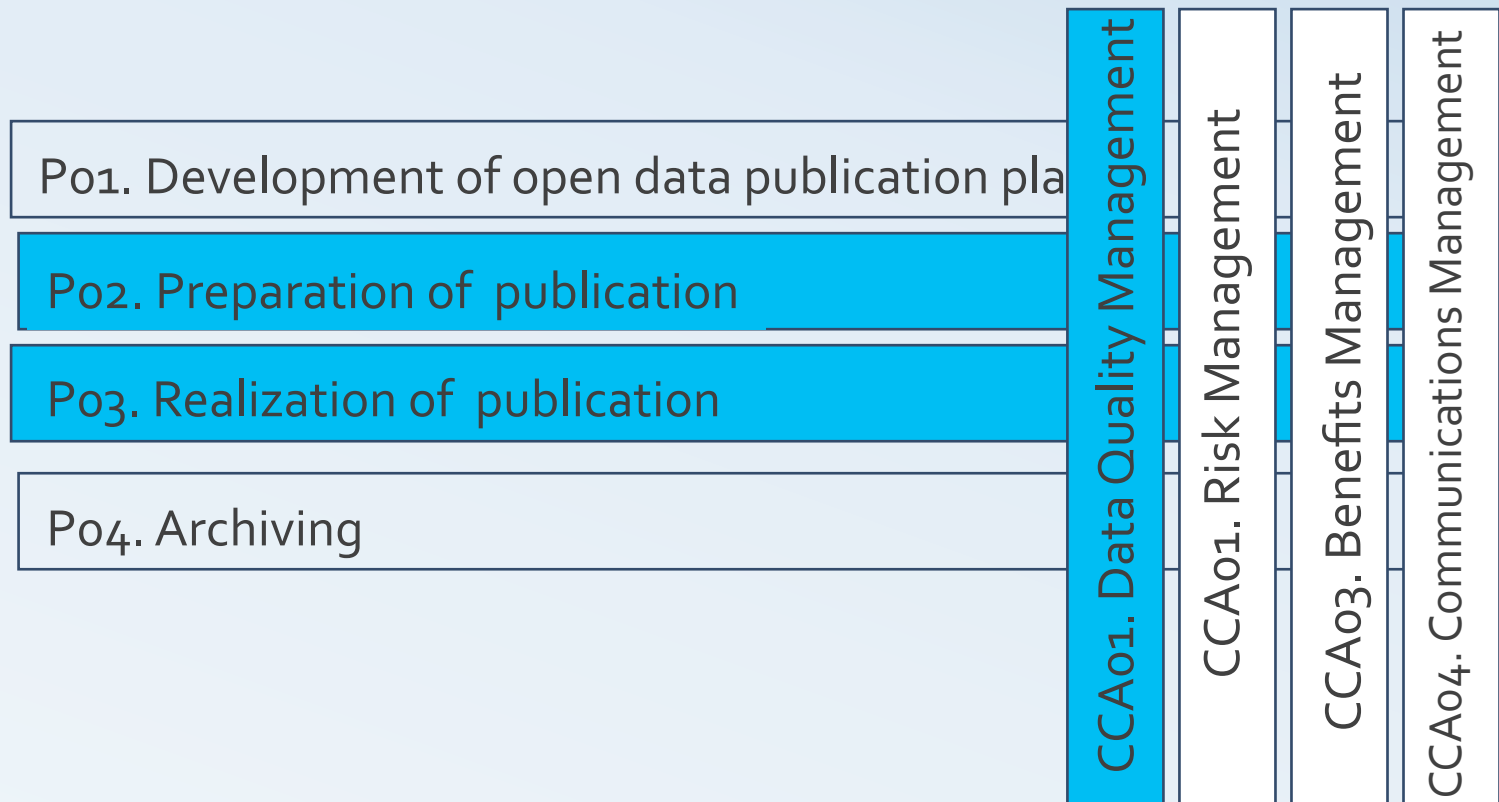


- **Publisher** - makes a dataset available to the public for download through a website.
- **Owner** - s/he holds rights to the dataset or s/he is legitimate to make decisions about the dataset.
- **Curator** - curates or maintains a dataset and its catalogue record (metadata); keeps the datasets accurate and up-to-date.
- **OD Catalogue Owner** - It is responsible for defining policies and rules governing cataloguing and its use.
- **OD Coordinator** - is responsible for coordination and management of the open data related activities of the organization.
- **OD Catalogue Publisher** - makes the data catalogue available to the potential users. OD Catalogue Publisher is responsible for operating the data catalog and s/he ensures maintenance of the underlying IT infrastructure.
- **IT Professional** - S/he provides support to other roles, develops and tests the ETL procedures and performs the transformation of the data into the target data formats.
- **Data Quality Manager** is in charge of supervising all data quality components and the data quality lifecycle.
- **Data Quality Expert** is in charge of analyze, apply and in case create the ETL components related to the data quality.
- **Legal Expert** - a person with skills and knowledge in the domain of law and legislation; provides his or her expertise about licencing of open datasets

Structure of the Methodology



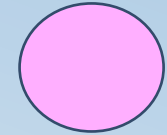
Addressed in more depth in COMSODE and in the following



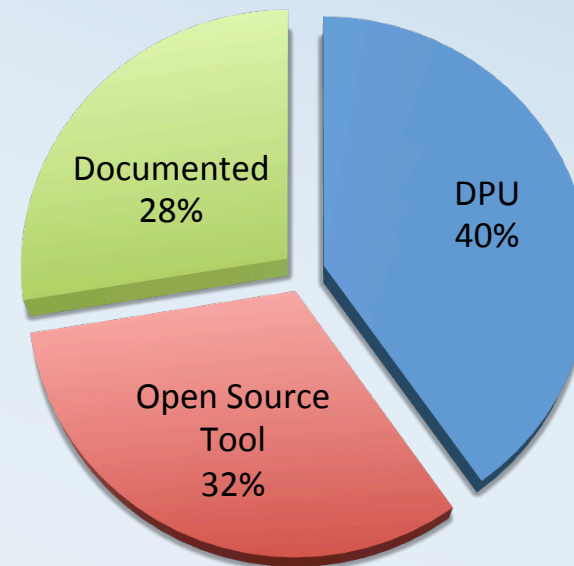
Figures about Techniques, DPU's, Phases,
Data Quality Dimensions, DQ SubPhases

Anatomy of COMSODE

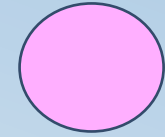
Figures about COMSODE Techniques



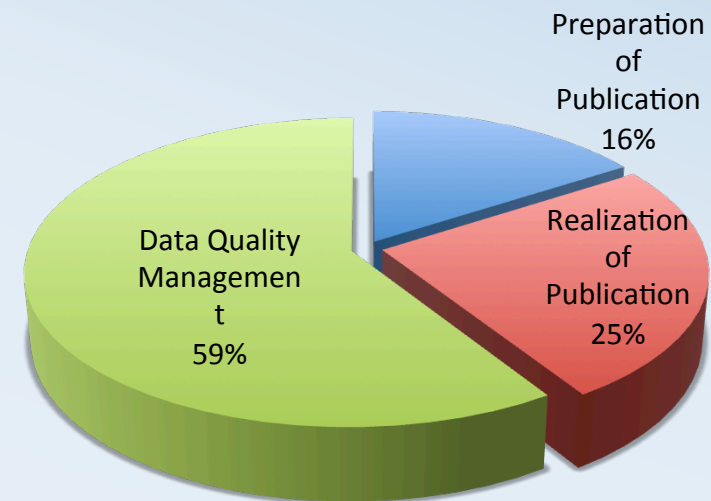
Type of technique	#
DPU	35
Open Source Tool	28
Documented	24



Figures about Techniques for Phases & DQ Management



Phase + DQM / Technique	#
Preparation of Publication	14
Realization of Publication	22
Data Quality Management	52

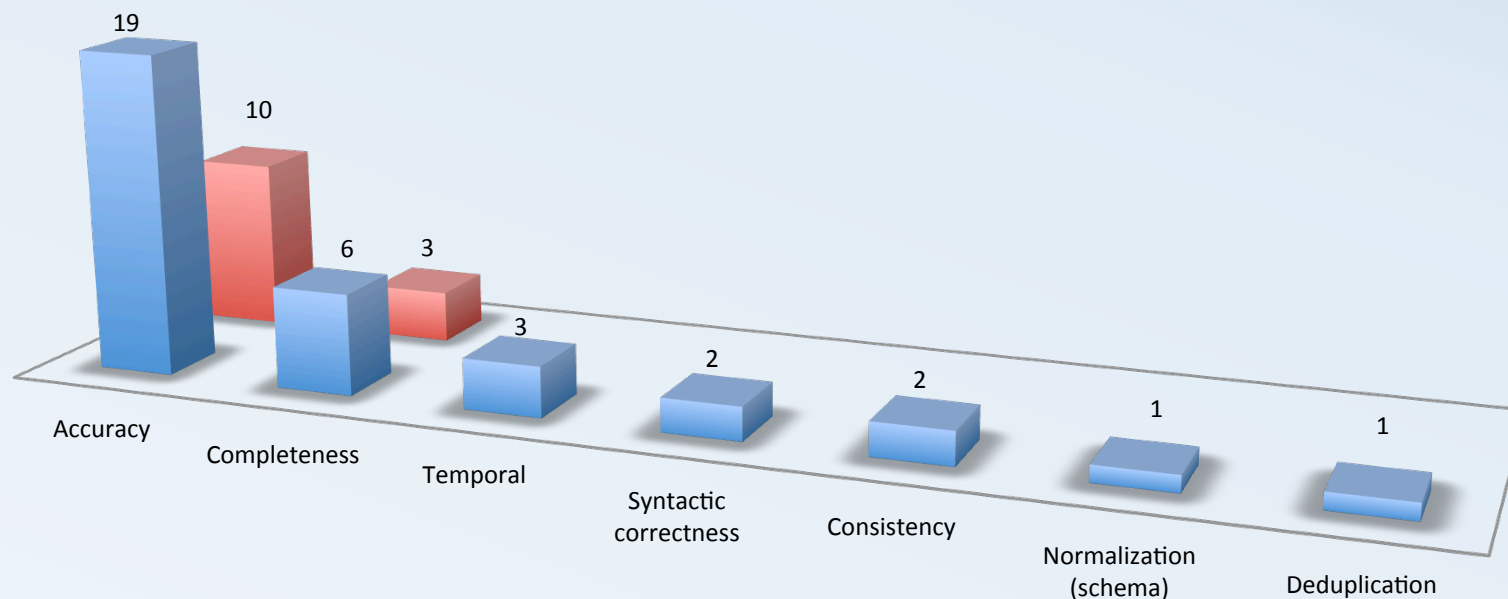


Figures about COMSODE

DQ Dimensions, DPUs and DQM Assessment vs Improvement



■ Assessment ■ Improvement



Data Quality Dimensions considered for the different data set types



Dimension ↓ / Format →	Metadata	CSV	Relational	Html	XML	RDF
Accuracy		- Syntactic accuracy - Semantic accuracy	-	-	-	Syntactic accuracy
Completeness	A measure for each metadata instance - Global weighted measure	- Object compl. - Tuple compl. - Attribute compl. - Table compl.	-	-	-	-
Consistency		Functional dependencies	-	-	-	-
Time related		Currency	-	-	-	- Age - Currency - Timeliness
Schema dimensions	-	-	Boyce Codd Normal Form	-	-Valid -Well formed	



Figures about COMSODE **DPU**s for **Step Types** in Preparation and Realization of Publication

Step	#
Extraction	4
Transformation	12
Schema design	2
Load	5
Anonymization	8
Integration	3

Figures about COMSODE

Techniques vs Data source format in input/output



DS format in input → /output ↓	CSV/Rel	CSV	Rel	RDF	Metad.	XML	All
CSV/Relational	38						
CSV			1	2			
Relational			6	1			
RDF		2		19			
Metadata					4		
XML				1		1	1
All				2			7

Figures about COMSODE

Techniques vs Data source format in input/output



DS format in input → /output ↓	CSV/Rel	CSV	Rel	RDF	Metad.	XML	All
CSV/Relational	38						
CSV			1	2			
Relational			6	1			
RDF		2		19			
Metadata					4		
XML				1		1	1
All				2			7

Two methodological cases



- COMSODE for a **single dataset**
- COMSODE for **multiple datasets**

Contents

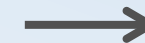
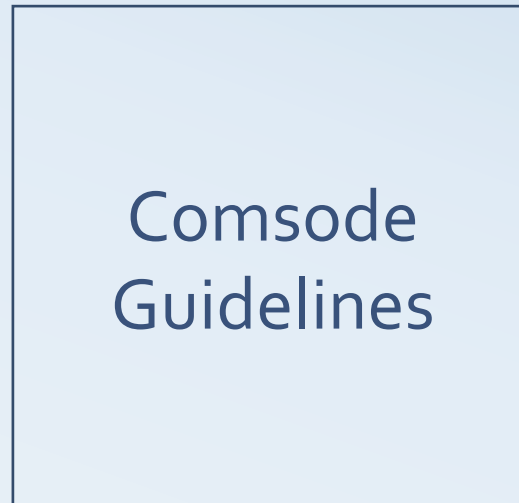
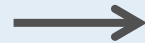


1. Introduction to Open Data and the COMSODE Publication Platform
2. Usage of COMSODE Techniques
3. The COMSODE Methodology: generalities
4. **Methodology for one dataset**
5. Methodology for multiple datasets
6. Social value of open data

Comsode for a single dataset



- One Data set
- Techniques,
- DPUs,
- Open source tools
- Doc. Techniques
- ODN platform services

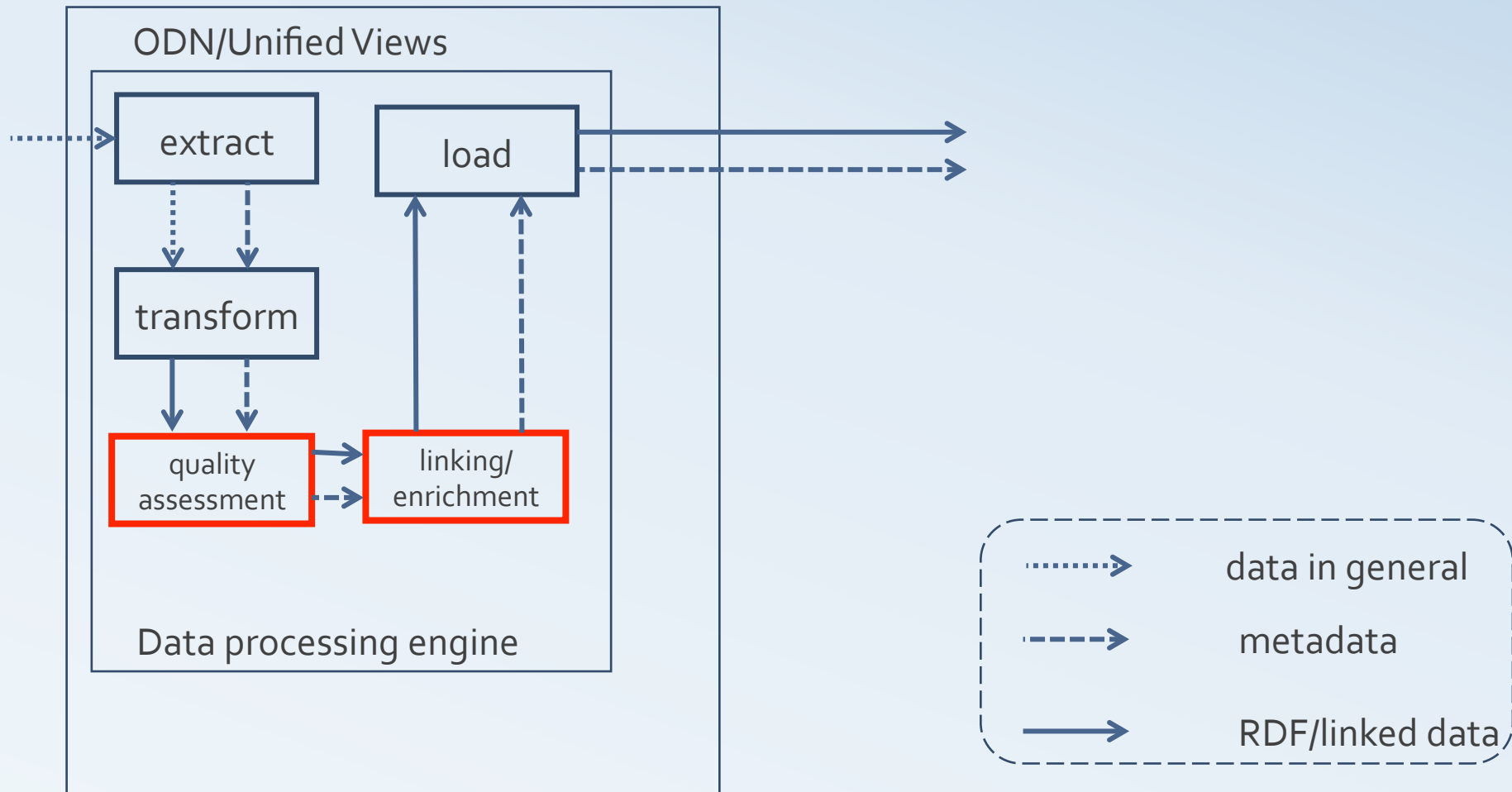


- Methodological path made of**
- Pipelines and
 - Steps using other techniques

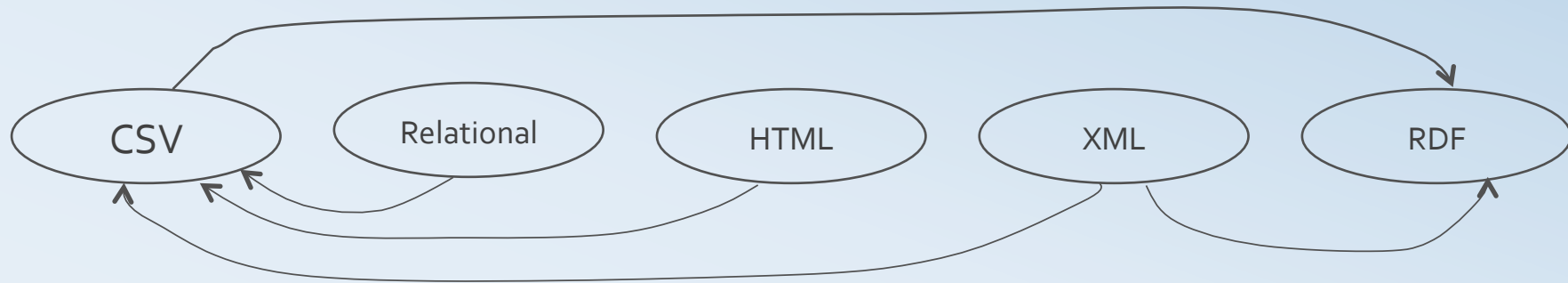


Domain Knowledge

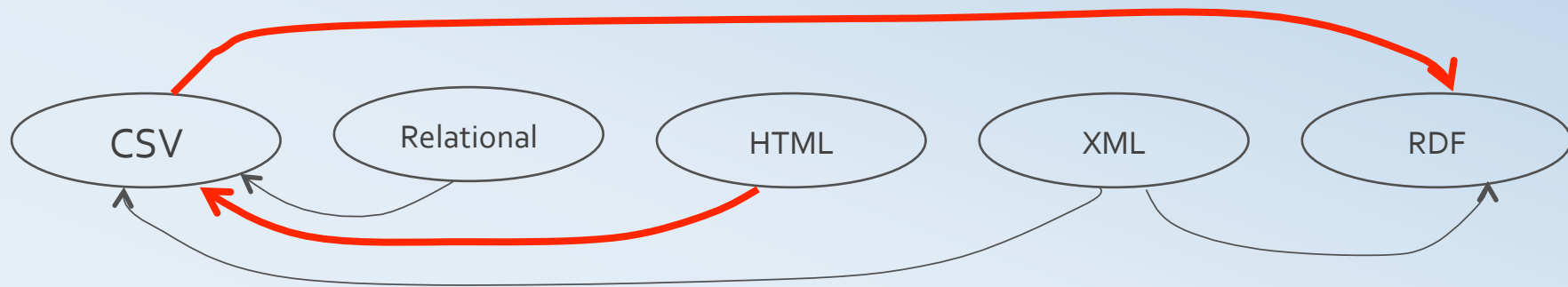
Single data sets



Types of data representations in input to the Comsode life cycle and suggested model transformations

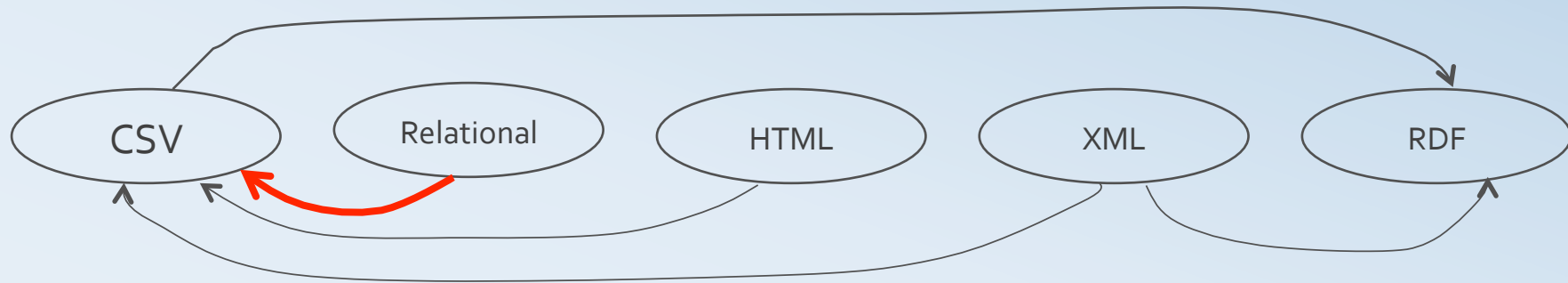


Types of data representations in input to the Comsode life cycle and suggested model transformations



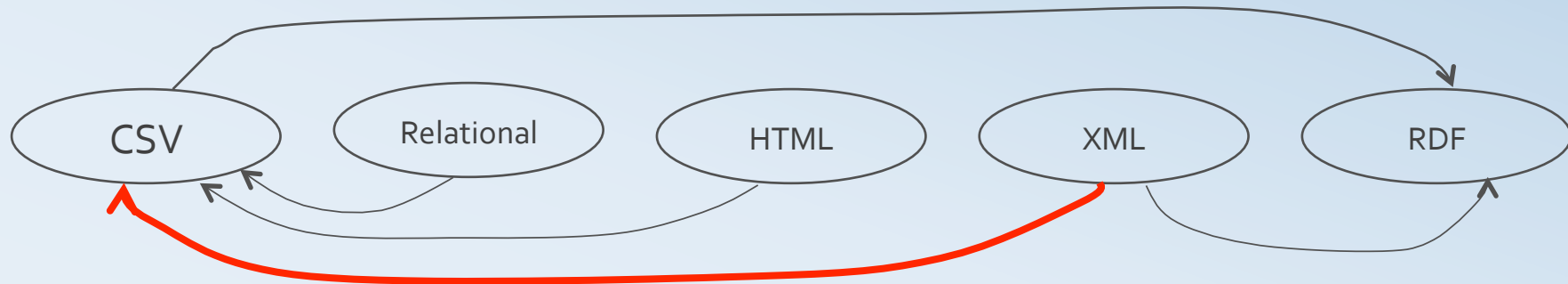
- If the dataset in input is **Html**, it is preferable to move to CSV, and at end to RDF

Types of data representations in input to the Comsode life cycle and suggested model transformations



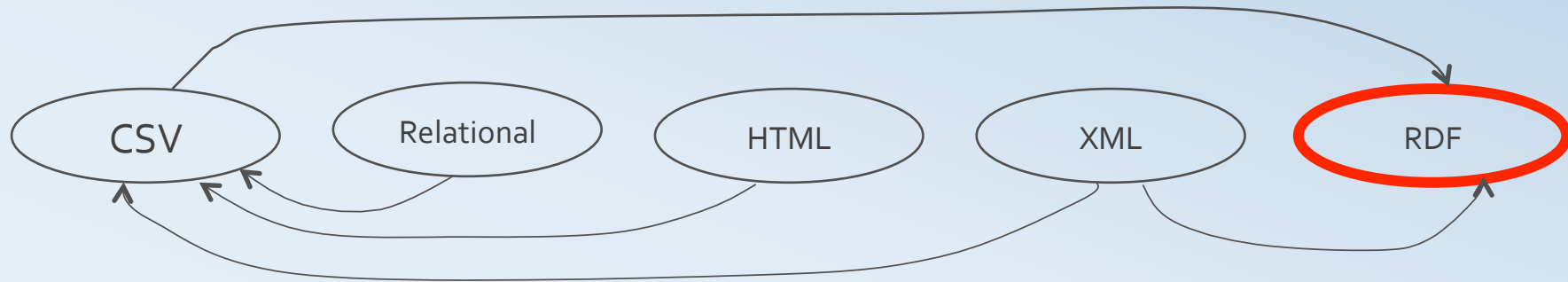
- If the dataset in input is **Html**, it is preferable to move to CSV, and at end to RDF
- If the dataset is **Relational**, you may keep Relational, but be careful since it is more complex than CSV

Types of data representations in input to the Comsode life cycle and suggested model transformations



- If the dataset in input is **Html**, it is preferable to move to CSV, and at end to RDF
- If the dataset is **Relational**, you may keep Relational, but be careful since it is more complex than CSV
- If you usually work with **hierarchical data**, and use **XML**, you may keep XML, but be careful since it is more complex, at end transform into CSV or RDF

Types of data representations in input to the Comsode life cycle and suggested model transformations



- If the dataset in input is **Html**, it is preferable to move to CSV, and at end to RDF
- If the dataset is **Relational**, you may keep Relational, but be careful since it is more complex than CSV
- If you usually work with **hierarchical data**, and use **XML**, you may keep XML, but be careful since it is more complex, at end transform into CSV or RDF
- If the dataset is **RDF**, perform the whole life cycle in RDF

We first see...

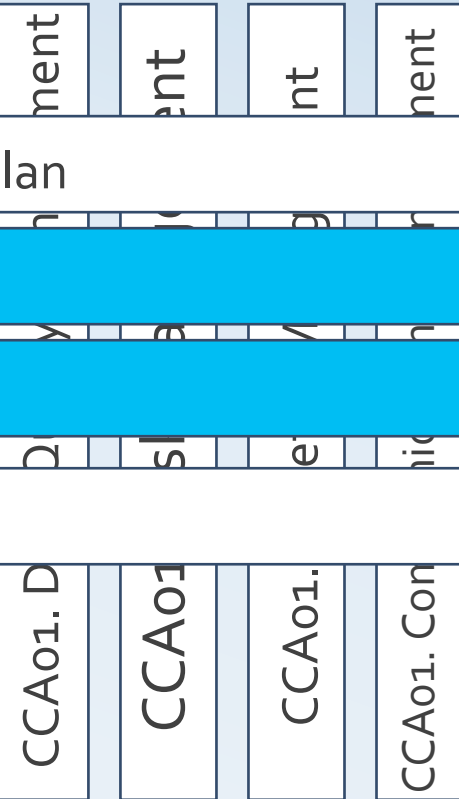


Po1. Development of open data publication plan

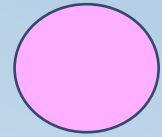
Po2. Preparation of publication

Po2. Realization of publication

Po2. Archiving

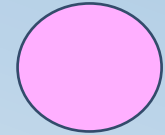


COMSODE for a single dataset: main activities (to be arranged in pipelines)



1. **Definition of the catalogue record schema** and the target data catalogues
2. **Designing data schemas and ontologies** for datasets published as (3*, 4* and) **5* data**
3. **Anonymization** of the dataset
4. **Designing extractors, transformers, and loaders**

More on 1. Definition of the catalogue record schema: specify metadata



RDF vocabularies such as **DCAT** and **VoiD** (and related DPUs) are used for the specification of **metadata** used for dataset description

Example for VoID

a. General dataset metadata

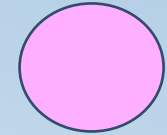
- Content
- Technical features
- Provenance
- History
- Licensing

b. Access metadata

c. Structural metadata

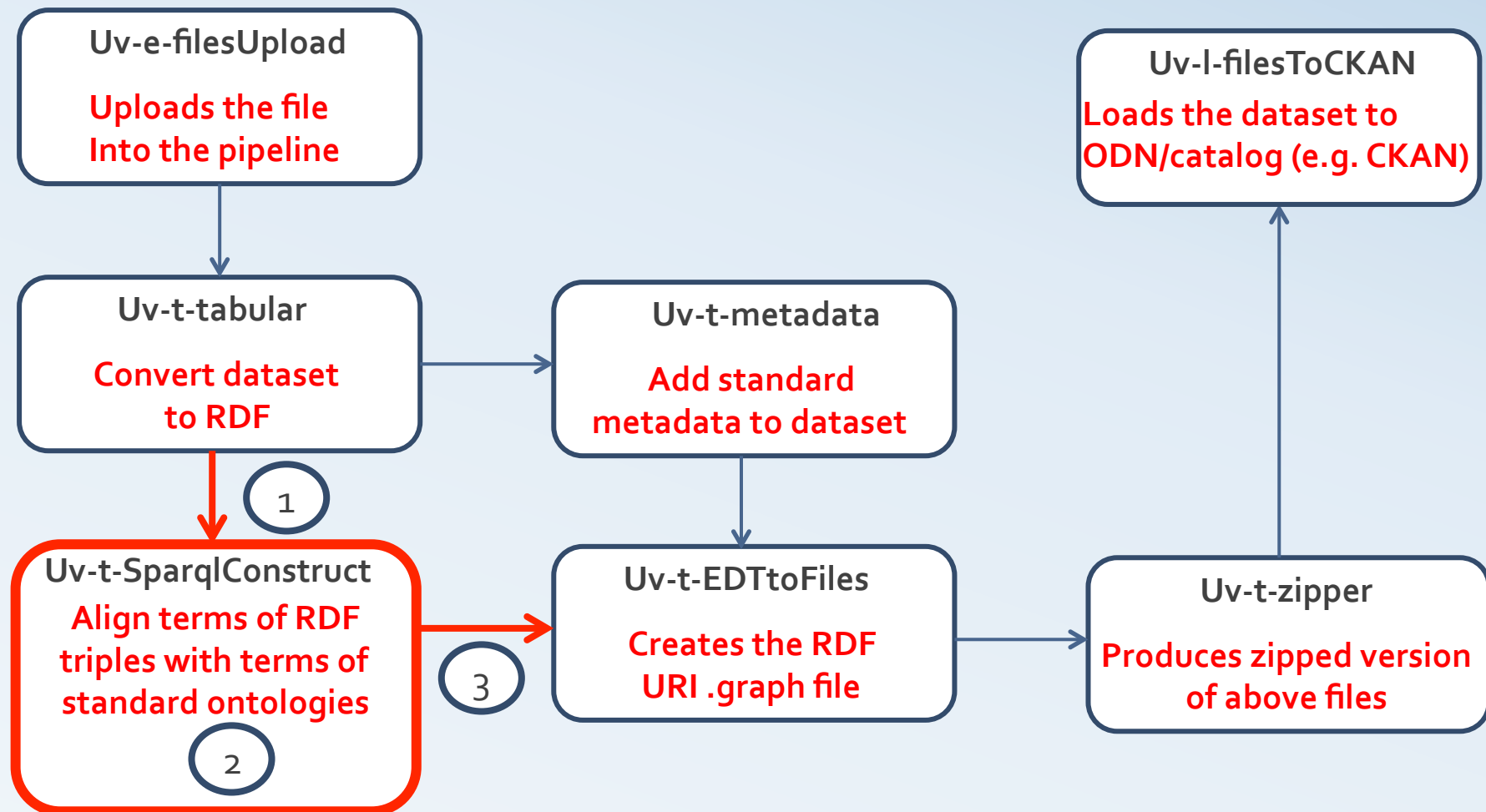
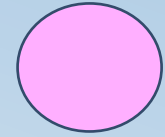
- Vocabularies used in the dataset
- Statistics about the size of the dataset
- Relevant partitions

More on 2. **Designing data schemas and ontologies for datasets published as 5* data** (represented in RDF)

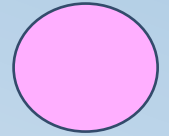


- A **data schema of a 5* dataset** is specified as a **vocabulary or ontology in one RDF-compatible language**, such as **RDFS and OWL2** that define the meaning of:
 - Classes
 - Datatypes and
 - Propertiesused to define the types of resources and literals, and the predicates.
- Users that **do not have** a strong background are recommended to use RDFS or a **small subset** of OWL2
- In the definition of a data schema it is suggested to **reuse existing vocabularies/ontologies**
- A **wide set** of catalogues, search engines, taxonomies and concept schemes is suggested for annotating data instances

The **SparqlConstruct** in action



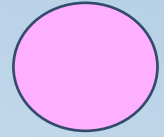
2. **Designing data schemas and ontologies** for datasets published as 5* data (represented in RDF)



Two reuse strategies

- **Maximum reuse strategy**, where the publisher uses the most popular vocabularies/ontologies
→ **maximum interoperability** for future integration with other datasets
- **Minimum reuse strategy**, where the publisher minimizes the reuse of external ontologies in order to strictly adhere to the specific domain nomenclature / dialect.
→ **maximum adherence** to the target domain

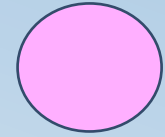
2. **Designing data schemas and ontologies** for datasets published as 5* data (represented in RDF)



Two methodologies

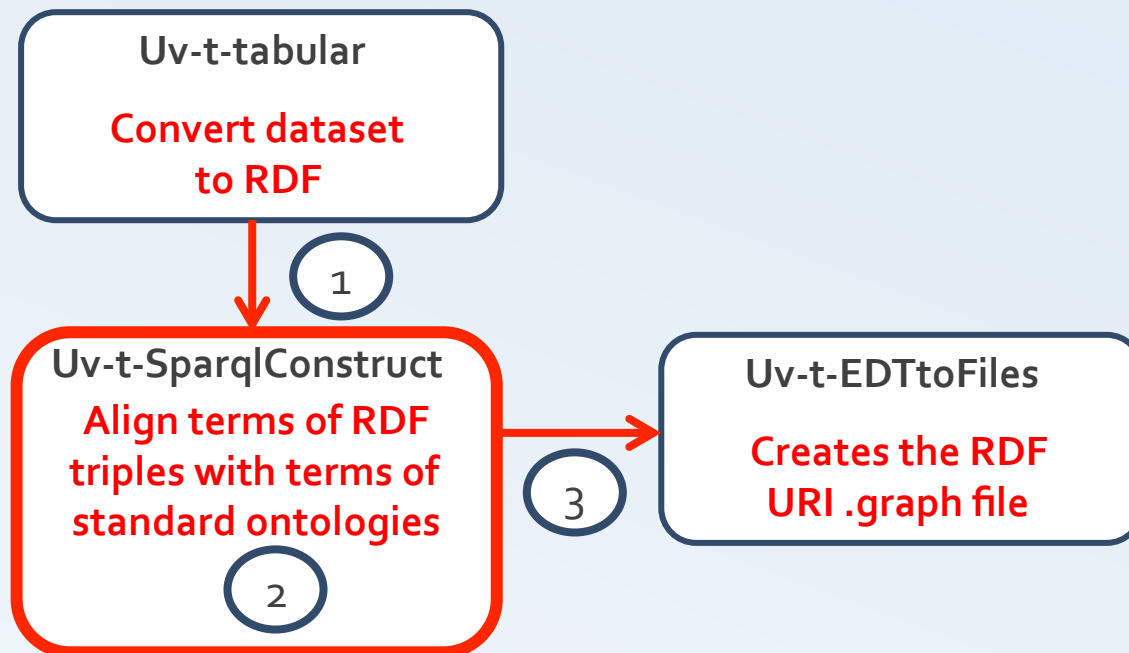
- **Top-down methodology** (ontology first, high-quality schema, higher cost), where the publisher
 - first defines the ontology and then
 - uses the published ontology to define a schema for the dataset;
- **Bottom-up methodology** (mapping first, low-quality schema, low cost), where the publisher
 - defines the schema at the same time when he/she transforms the data in RDF,
 - Then better specifies the ontology/vocabulary of the schema by subsequent transformations, mappings and other operations.

3. **Anonymization** of the dataset three **approaches**

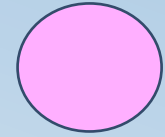


- **Projection**, where **particular attributes** with protected data **are removed** from the dataset, e.g., in case of tabular files, it means removing a column or columns.
- **Aggregation** approaches aim to **data perturbation** (e.g. by means of generalization, suppression or permutation, and randomization of values), in order to minimize the chances of identifying individual records while maximizing the accuracy of data querying.
- **Link removal** or reduction **removes the external links** that may hinder data-privacy before publishing the dataset.

The Tabular & SparqlConstruct DPU_s in action



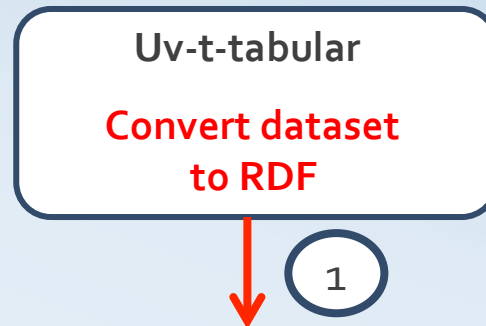
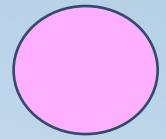
1. The T-Tabular DPU: transformation from CSV to RDF



T-Tabular
Transforms
a CSV dataset
into an RDF dataset

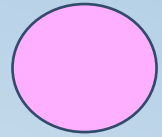
- Each **row** in the CSV dataset is converted into an RDF **resource**, identified by an URI.
- Each **attribute** in the CSV file is treated as an RDF **property**, so that the DPU outputs a triple for each attribute of each record.

2. Result of **transforming the csv into rdf** with t-tabular DPU



```
<http://demo.comsode.eu/ny-hospitals/2>
  <http://demo.comsode.eu/ny-hospitals/provider_id> "330003"@en ;
  <http://demo.comsode.eu/ny-hospitals/hospital_name> "ALBANY MEMORIAL HOSPITAL"@en ;
  <http://demo.comsode.eu/ny-hospitals/address> "600 NORTHERN BOULEVARD"@en ;
  <http://demo.comsode.eu/ny-hospitals/city> "ALBANY"@en ;
  <http://demo.comsode.eu/ny-hospitals/state> "NY"@en ;
  <http://demo.comsode.eu/ny-hospitals/zip_code> "12204"@en ;
  <http://demo.comsode.eu/ny-hospitals/measure_name> "Rate of complications for hip/
knee replacement patients"@en ;
  <http://demo.comsode.eu/ny-hospitals/measure_id> "COMP_HIP_KNEE"@en ;
  <http://linked.opendata.cz/ontology/odcs/tabular/row> "2"^^<http://www.w3.org/
2001/XMLSchema#int> ;
  a <http://unifiedviews.eu/ontology/t-tabular/Row> .
```

3. Sparql Construct query to map the dataset to **schema.org**



PREFIX c:<http://demo.comsode.eu/ny-hospitals/>

PREFIX s:<http://schema.org/>

CONSTRUCT{

?hospital a s:MedicalEntity.

?hospital a s:Organization.

_:mc a s:MedicalCode.

?hospital s:code _:mc.

_:mc s:codeValue ?hospital_id.

?hospital s:name ?hospital_name.

_:p a s:PostalAddress.

?hospital s:location _:p.

_:p s:streetAddress ?hospital_address.

_:p s:postalCode ?hospital_zipcode.

_:p s:addressLocality ?hospital_city.

_:p s:addressRegion ?hospital_country.

_:ms a s:MedicalStudy.

?hospital s:study _:ms.

_:ms s:outcome ?hospital_measurename.

_:ms s:alternatename ?hospital_measureid.

}

Uv-t-SparqlConstruct

Align terms of RDF
triples with terms of
standard ontologies

2

WHERE{

?hospital c:hospital_name ?hospital_name.

?hospital c:address ?hospital_address.

?hospital c:provider_id ?hospital_id.

?hospital c:city ?hospital_city.

?hospital c:state ?hospital_country.

?hospital c:zip_code ?hospital_zipcode.

?hospital c:measure_name ?

hospital_measurename.

?hospital c:measure_id ?hospital_measureid

}

4. Result of **alignment** of the hospital dataset with **schema.org**

Uv-t-SparqlConstruct

Align terms of RDF
triples with terms of
standard ontologies

3

<http://demo.comsode.eu/ny-hospitals/1193> a <http://schema.org/MedicalEntity> , <http://schema.org/Organization> ;

<http://schema.org/code> _:node1a1iu8btlx3578 ;

<http://schema.org/name> "CARTHAGE AREA HOSPITAL, INC"@en ;

<http://schema.org/location> _:node1a1iu8btlx3579 ;

<http://schema.org/study> _:node1a1iu8btlx3580 .

_:node1a1iu8btlx3578 a <http://schema.org/MedicalCode> ;

<http://schema.org/codeValue> "331318"@en .

_:node1a1iu8btlx3579 a <http://schema.org/PostalAddress> ;

<http://schema.org/streetAddress> "1001 WEST STREET"@en ;

<http://schema.org/postalCode> "13619"@en ;

<http://schema.org/addressLocality> "CARTHAGE"@en ;

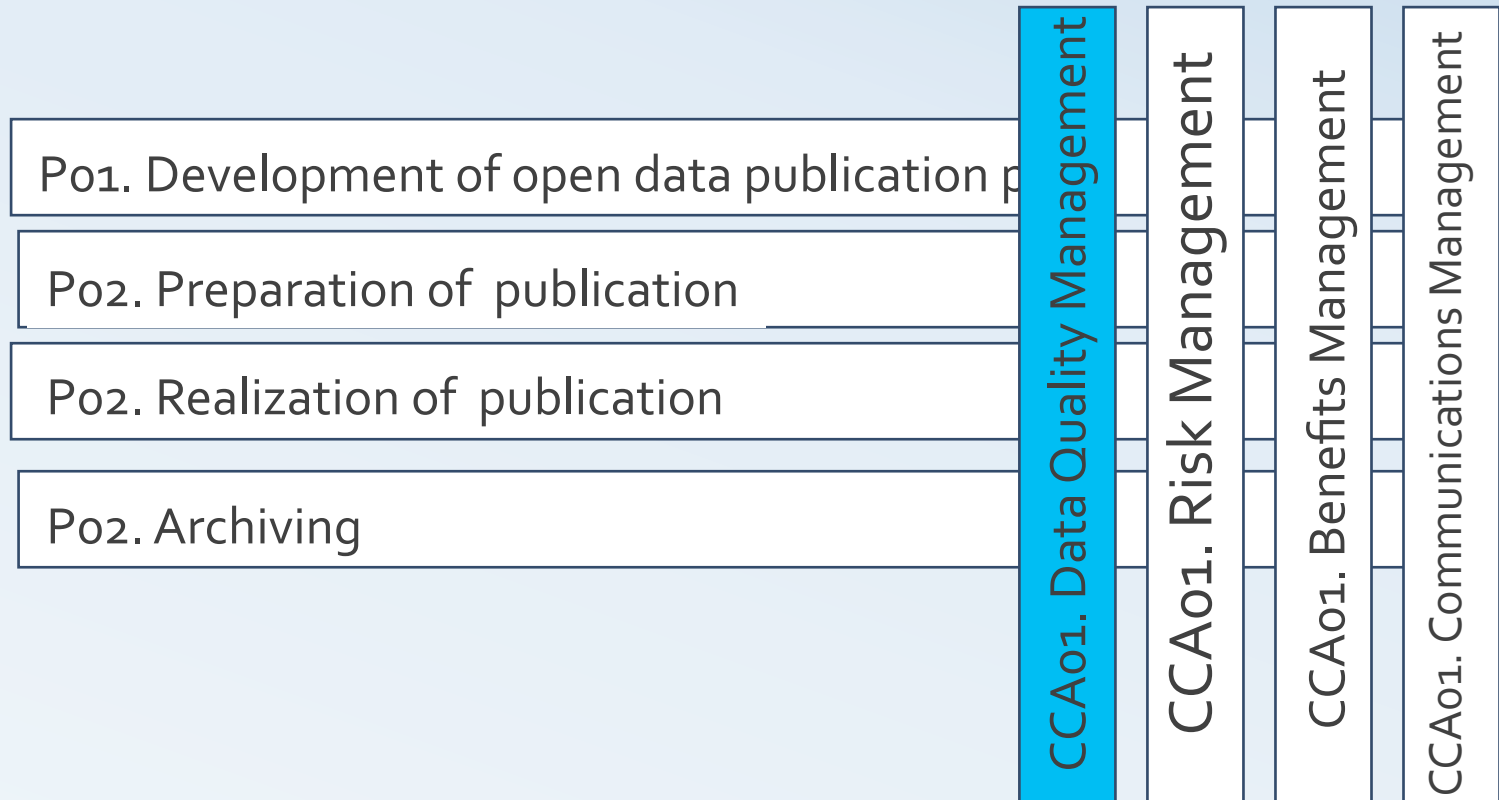
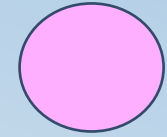
<http://schema.org/addressRegion> "NY"@en .

_:node1a1iu8btlx3580 a <http://schema.org/MedicalStudy> ;

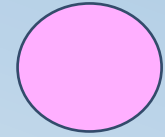
<http://schema.org/outcome> "Serious complications"@en ;

<http://schema.org/alternatename> "PSI_90_SAFETY"@en .

We move to...



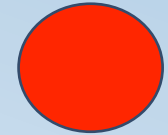
Data quality management



1. **REQUIREMENTS COLLECTION AND ANALYSIS** – Collect general requirements, achieve from users specific requirements and subjective dataset quality evaluation and possible causes of errors using the COMSODE questionnaire. Analyze requirements and select priority quality dimensions.
3. **QUANTITATIVE ASSESSMENT** - Perform quantitative assessment on selected quality dimensions, choosing suitable quality metrics and related techniques
4. **IMPROVEMENT** - Choose suitable improvement activities and related techniques. Apply chosen improvement techniques.

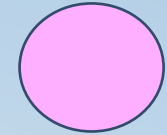
Assessment and improvement of quality

Examples of DPUs



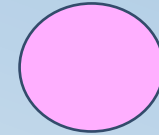
- **Q-ACC_1**: detection of *ill-typed literals*
- **Q-ACC_2**: detecting *syntax errors* using validators;
- **Q-ACC_7**: use of regular expressions to *identify date information*;
- **Q-C_1**: property completeness
- **Q-C_2**: dataset completeness
- **Q-C_5**: understandability of a resource
- **Q-C_1**: currency of a document, measured as the *age of the document*.

Example of Data quality Management: input table



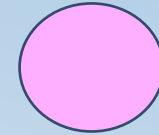
Tuple #	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name T.oponym	Number T.	City	Zip Code	Country
1	Miroslav	Konecny	1978	7	0	Street	St. John	49	Prague	412776	null
2	Martin	Necasky	1975	7	3	Sq.	Vienna	null	Bratislava	101278	Slovensko
3	Miroslav	Konecny	null	null	null	Str.	Saint Jon	49	Prague	412776	null
4	Crlo	Btini	1949	June	7	Street	Dessiè	15	Roma	00198	Italia
5	Miroslav	Knecy	1978	7	null	Sq.	Budapest	23	Wien	null	Austria
6	Anisa	Rula	1982	September	7	Via	Sesto	null	Milano	20...	Ital
7	Anita	Rula	1982	9	7	Via	Seto	23	Milan	null	Italy
8	Anna	Rla	null	null	null	Via	Sarca	336	Milano	null	Italy
9	Carlo	Batini	1949	6	7	V.	Beato Angelico	23	Milan	20133	Italy
10	Carla	Botni	1949	June	7	Av.	Charles	null	Prague	412733	null
11	Marta	Necasky	1976	11	3	null	null	null	Bratislava	112234	Slov.

Output of the **Standardization Step**



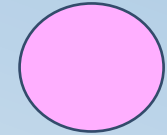
Tuple #	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name T.oponym	Number T.	City	Zip Code	Country
1	Miroslav	Konecny	1978	7	0	<u>Street</u>	St. John	49	Prague	412776	<u>null</u>
2	Martin	Necasky	1975	7	3	<u>Square</u>	Vienna	null	Bratislava	101278	<u>Slovacchia</u>
3	Miroslav	Konecny	null	null	null	<u>Street</u>	Saint Jon	49	Prague	412776	<u>null</u>
4	Crlo	Botini	1949	6	7	<u>Street</u>	Dessiè	15	Roma	00198	<u>Italia</u>
5	Miroslav	Knecy	1978	7	null	<u>Square</u>	Budapest	23	Wien	null	<u>Austria</u>
6	Anisa	Rula	1982	9	7	<u>Street</u>	Sesto	null	Milano	20...	<u>Italy</u>
7	Anita	Rula	1982	9	7	<u>Street</u>	Seto	23	Milan	null	<u>Italy</u>
8	Anna	Rla	null	null	null	<u>Street</u>	Sarca	336	Milano	null	<u>Italy</u>
9	Carlo	Batini	1949	6	7	<u>Street</u>	Beato Angelico	23	Milano	20133	<u>Italy</u>
10	Carla	Botni	1949	6	7	<u>Avenue</u>	Charles	null	Prague	412733	<u>null</u>
11	Marta	Necasky	1976	11	3	<u>null</u>	null	null	Bratislava	112234	<u>Slovacchia</u>

First improvement based on syntactic accuracy



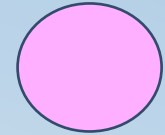
Tuple #	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name T.oponym	Number T.	City	Zip Code	Country
1	Miroslav	Konecny	1978	7	10	Street	Saint John	49	Prague	412776	null
2	Martin	Necasky	1975	7	8	Square	Wien	null	Bratislava	101278	Slovacchia
3	Miroslav	Konecny	null	null	null	Street	Saint <u>John</u>	49	Prague	412776	null
4	Carlo	Btini	1949	6	7	Street	Dessiè	15	Roma	00198	Italia
5	Miroslav	Knecy	1978	7	null	Square	Budapest	23	Wien	null	Austria
6	Anisa	Rula	1982	9	7	Street	Sesto	null	Milano	20...	<u>Italy</u>
7	Anita	Rula	1982	9	7	Street	<u>Sesto</u>	23	<u>Milano</u>	null	Italy
8	Anna	Rla	null	null	null	Street	Sarca	336	Milano	null	Italy
9	Carlo	Batini	1949	6	7	Street	Beato Angelico	23	Milano	20133	Italy
10	Carla	Batni	1949	6	7	Avenue	Charles	null	Prague	412733	null
11	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slovacchia

Improvement based on completeness



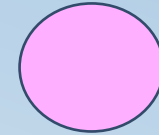
Tuple #	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name T.oponym	Number T.	City	Zip Code	Country
1	Miroslav	Konecny	1978	7	10	Street	Saint John	49	Prague	412776	<u>Czech</u>
2	Martin	Necasky	1975	7	8	Square	Wien	null	Bratislava	101278	Slovacchia
3	Miroslav	Konecny	null	null	null	Street	Saint <u>John</u>	49	Prague	412776	<u>Czech</u>
4	Carlo	Btini	1949	6	7	Street	Dessiè	15	Roma	00198	Italy
5	Miroslav	Knecy	1978	7	null	Square	Budapest	23	Wien	12345	Austria
6	Anisa	Rula	1982	9	7	Street	Sesto	null	Milano	<u>20127</u>	<u>Italy</u>
7	Anita	Rula	1982	9	7	Street	<u>Sesto</u>	23	<u>Milano</u>	<u>20127</u>	Italy
8	Anna	Rla	null	null	null	Street	Sarca	336	Milano	<u>20126</u>	Italy
9	Carlo	Batini	1949	6	7	Street	Beato Angelico	23	Milano	20133	Italy
10	Carla	Batni	1949	6	7	Avenue	Charles	null	Prague	412733	<u>Czech</u>
11	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slovacchia

Consistency check



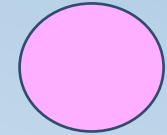
Tuple #	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name T.oponym	Number T.	City	Zip Code	Country
1	Miroslav	Konecny	1978	7	10	Street	Saint John	49	Prague	412776	Czech
2	Martin	Necasky	1975	7	8	Square	Wien	null	Bratislava	101278	Slovacchia
3	Miroslav	Konecny	null	null	null	Street	Saint John	49	Prague	412776	Czech
4	Carlo	Btini	1949	6	7	Street	Dessiè	15	Roma	<u>00199</u>	Italy
5	Miroslav	Knecy	1978	7	null	Square	Budapest	23	Wien	k2345	Austria
6	Anisa	Rula	1982	9	7	Street	Sesto	null	Milano	20127	Italy
7	Anita	Rula	1982	9	7	Street	Sesto	23	Milano	20127	Italy
8	Anna	Rla	null	null	null	Street	Sarca	336	Milano	20126	Italy
9	Carlo	Batini	1949	6	7	Street	Beato Angelico	23	Milano	20133	Italy
10	Carla	Batni	1949	6	7	Avenue	Charles	null	Prague	412733	Czech
11	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slovacchia

Candidate tuples for deduplication



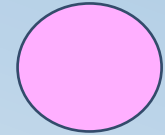
Tuple #	Candidate tuples	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name T.oponym	Number T.	City	Zip Code	Country
1	M1	Miroslav	Konecny	1978	7	10	Street	Saint John	49	Prague	412776	Czech
2	M2	Martin	Necasky	1975	7	8	Square	Wien	null	Bratislava	101278	Slovacchia
3	M1	Miroslav	Konecny	null	null	null	Street	Saint John	49	Prague	412776	Czech
4	PM1	Carlo	Btini	1949	6	7	<u>Street</u>	<u>Dessiè</u>	15	<u>Roma</u>	00199	Italy
5	M2	Miroslav	Knecy	1978	7	null	Square	Budapest	23	Wien	k2345	Austria
6	M3	Anisa	Rula	1982	9	7	Street	Sesto	null	Milano	20127	Italy
7	M3	Anita	Rula	1982	9	7	Street	Sesto	23	Milano	20127	Italy
8	NM1	Anna	Rla	null	null	null	Street	Sarca	336	Milano	20126	Italy
9	PM1	Carlo	Batini	1949	6	7	<u>Street</u>	<u>Beato Angelico</u>	23	<u>Milano</u>	20133	Italy
10	NM2	Carla	Batni	1949	6	7	Avenue	Charles	null	Prague	412733	Czech
11	NM3	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slovacchia

Result of **currency check**



Tuple #	Candidate tuples	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name T.oponym	Number T.	City	Zip Code	Country
1	M1	Miroslav	Konecny	1978	7	10	Street	Saint John	49	Prague	412776	Czech
2	M2	Martin	Necasky	1975	7	8	Square	Wien	null	Bratislava	101278	Slovacchia
3	M1	Miroslav	Konecny	null	null	null	Street	Saint John	49	Prague	412776	Czech
4	PM1	Carlo	Btini	1949	6	7	<u>Street</u>	<u>Beato Angelico</u>	<u>23</u>	<u>Milano</u>	20133	Italy
5	M2	Miroslav	Knecy	1978	7	null	Square	Budapest	23	wien	k2345	Austria
6	M3	Anisa	Rula	1982	9	7	Street	Sesto	null	Milano	20127	Italy
7	M3	Anita	Rula	1982	9	7	Street	Sesto	23	Milano	20127	Italy
8	NM1	Anna	Rla	null	null	null	Street	Sarca	336	Milano	20126	Italy
9	PM1	Carlo	Batini	1949	6	7	Street	Beato Angelico	23	Milano	20133	Italy
10	NM2	Carla	Batni	1949	6	7	Avenue	Charles	null	Prague	412733	Czech
11	NM3	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slovacchia

Final table with **duplicates highlighted**



Tuple #	Candidate tuples	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name T.oponym	Number T.	City	Zip Code	Country
1	M1	Miroslav	Konecny	1978	7	10	Street	Saint John	49	Prague	412776	Czech
2	M2	Martin	Necasky	1975	7	8	Square	Wien	null	Bratislava	101278	Slovacchia
3	M1	Miroslav	Konecny	null	null	null	Street	Saint John	49	Prague	412776	Czech
4	M4	Carlo	Btini	1949	6	7	<u>Street</u>	<u>Dessie</u>	<u>15</u>	<u>Roma</u>	00199	Italia
5	M2	Miroslav	Knecy	1978	7	null	Square	Budapest	23	Wien	k2345	Austria
6	M3	Anisa	Rula	1982	9	7	Street	Sesto	null	Milano	20127	Italy
7	M3	Anita	Rula	1982	9	7	Street	Sesto	23	Milano	20127	Italy
8	NM1	Anna	Rla	null	null	null	Street	Sarca	336	Milano	20126	Italy
9	M4	Carlo	Batini	1949	6	7	<u>Street</u>	<u>Beato Angelico</u>	<u>23</u>	<u>Milano</u>	20133	Italy
10	NM2	Carla	Batni	1949	6	7	Avenue	Charles	null	Prague	412733	Czech
11	NM3	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slovacchia

Contents

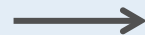


1. Introduction to Open Data and the COMSODE Publication Platform
2. Usage of COMSODE Techniques
3. The COMSODE Methodology: generalities
4. Methodology for one dataset
5. **Methodology for multiple datasets**
6. Social value of open data

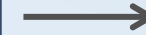
Comsode for multiple dataset



- CSV Datasets
- RDF/XML data set(s)
- Techniques,
- DPUs,
- Open source tools
- Doc. Techniques
- ODN platform services

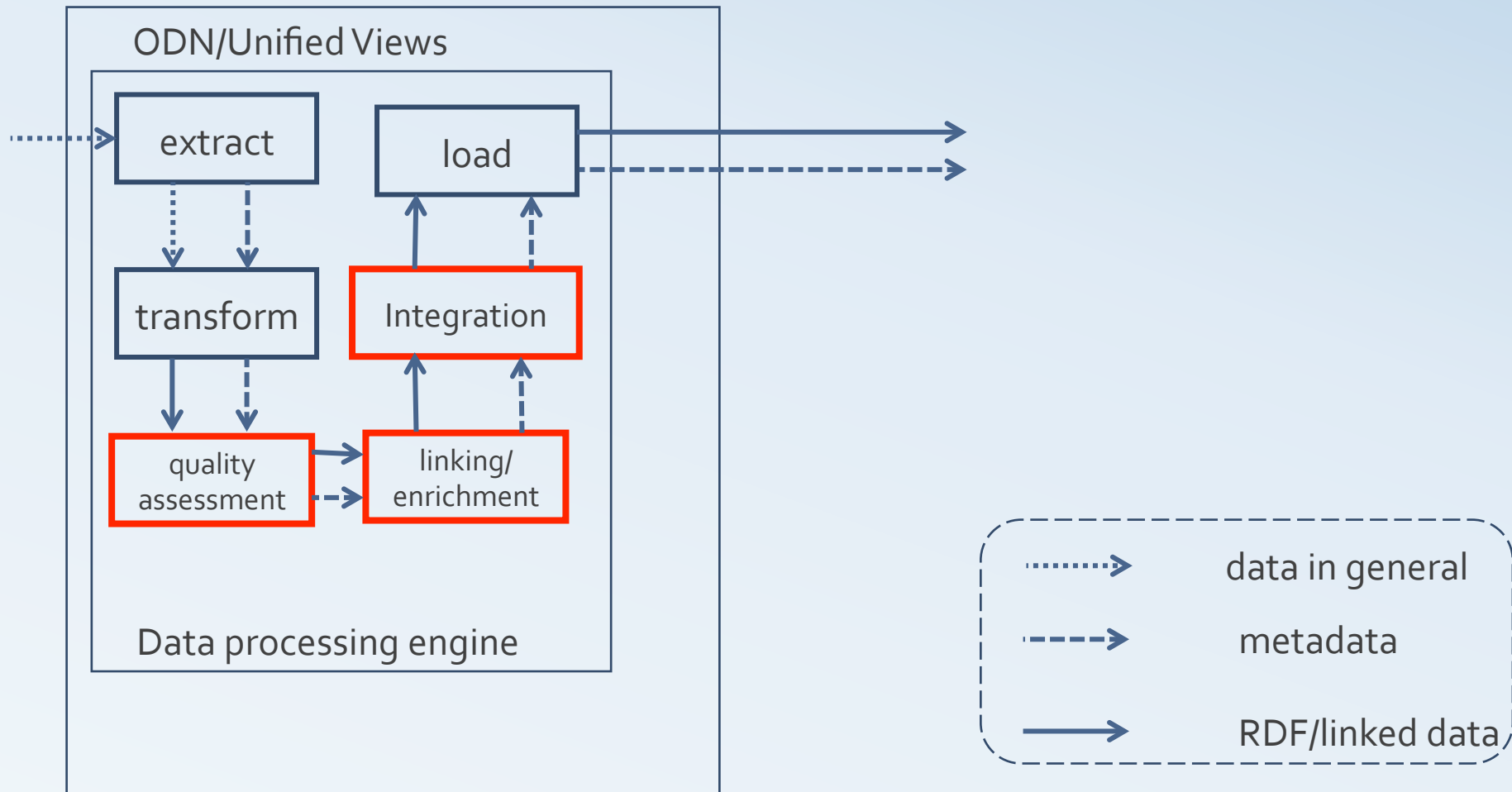


Comsode
Guidelines



- Methodological path made of**
- Pipelines and
 - Steps using other techniques

Multiple data sets

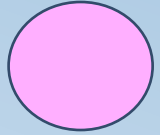


COMSODE for multiple datasets: new integration steps



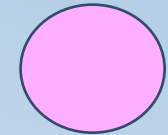
1. **License matching for compatibility assessment**
2. **Domain knowledge exploitation**
3. **Choice of the integration process**
4. **Pre-integration**
5. **Transformation**
6. **Matching**
7. **Local reconciliation and final integration**
8. **Maintenance**

Traditional integration process with relational + ER data + schema

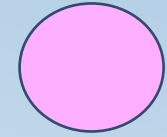


Kind of input → Activity ↓		Schema	Dataset
3. Choice of the integration process		Intra-model or inter-model integration	Matching directly or with external sources (match or link and match) / local or global integration
4. Pre-integration		Preparation of schema (includes standardization, normalization)	Preparation of data (includes standardization, cleansing)
6. Matching		Find correspondences & heterogeneities	Final matching / non matching pairs
7. Local reconciliation and final Integration		Solve heterogeneities & keep correspondences	Data fusion

Enriched integration process for CSV + RDF data + schema



Kind of input → Activity ↓	Metadata	Schema	Dataset
1, Licence matching for Compatibility assessment	Licence fields of metadata	-	-
2. Domain knowledge Exploitation	-	Local schema alignment with ontologies and vocabularies	Tabular annotation (if CSV) or local data alignment with external knowledge sources (if RDF)
3. Choice of the integration process	-	Intra-model or inter-model integration	Matching directly or with external sources (match or link and match) / local or global integration
4. Pre-integration		Preparation of schema (includes standardization, normalization)	Preparation of data (includes standardization, cleansing)
5. Transformation	-	Translation from one format to another	Transformation from one format to another
6. Matching	-	Find correspondences & heterogeneities	Final matching / non matching pairs
7. Local reconciliation and Final Integration	-	Solve heterogeneities & keep correspondences	Data fusion
8. Maintenance	-	Decides on rebuilding the process from scratch or incrementally	Decides on rebuilding the process from scratch or incrementally



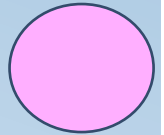
More on

1. Licence compatibility assessment

More on licence compatibility assessment

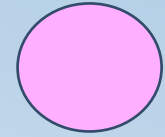
A **Licence** is composed of different **Models**,
Each model is defined as a set of **Elements** characterized by different levels of compatibility (see next slide)

Model name	Model definition	Element name	Element definition
Permission	Actions that may or may not be allowed or desired	Reproduction	making multiple copies
		Distribution	distribution, public display, and publicly performance
		Derivative Works	distribution of derivative works, i.e., works modified w.r.t. the original
		Sharing	permits commercial derivatives, but only non-commercial distribution
Requirement	Actions that may or may not be requested	Notice	copyright and license notices be kept intact
		Attribution	credit be given to copyright holder and/or author
		Attach Policy	attach the policy to the work
		Share Alike	derivative works be licensed under the same terms or compatible terms as the original work
		Source Code	source code (the preferred form for making modifications) must be provided when exercising some rights granted by the license
		Copyleft	derivative and combined works must be licensed under specified terms, similar to those on the original work
Prohibition	Actions that are asked not to be done	Commercial use	exercising rights for commercial purposes
		High income Nation Use	use in a non-developing country



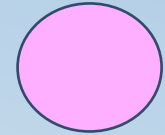
Compatibility rules

for the Permission element



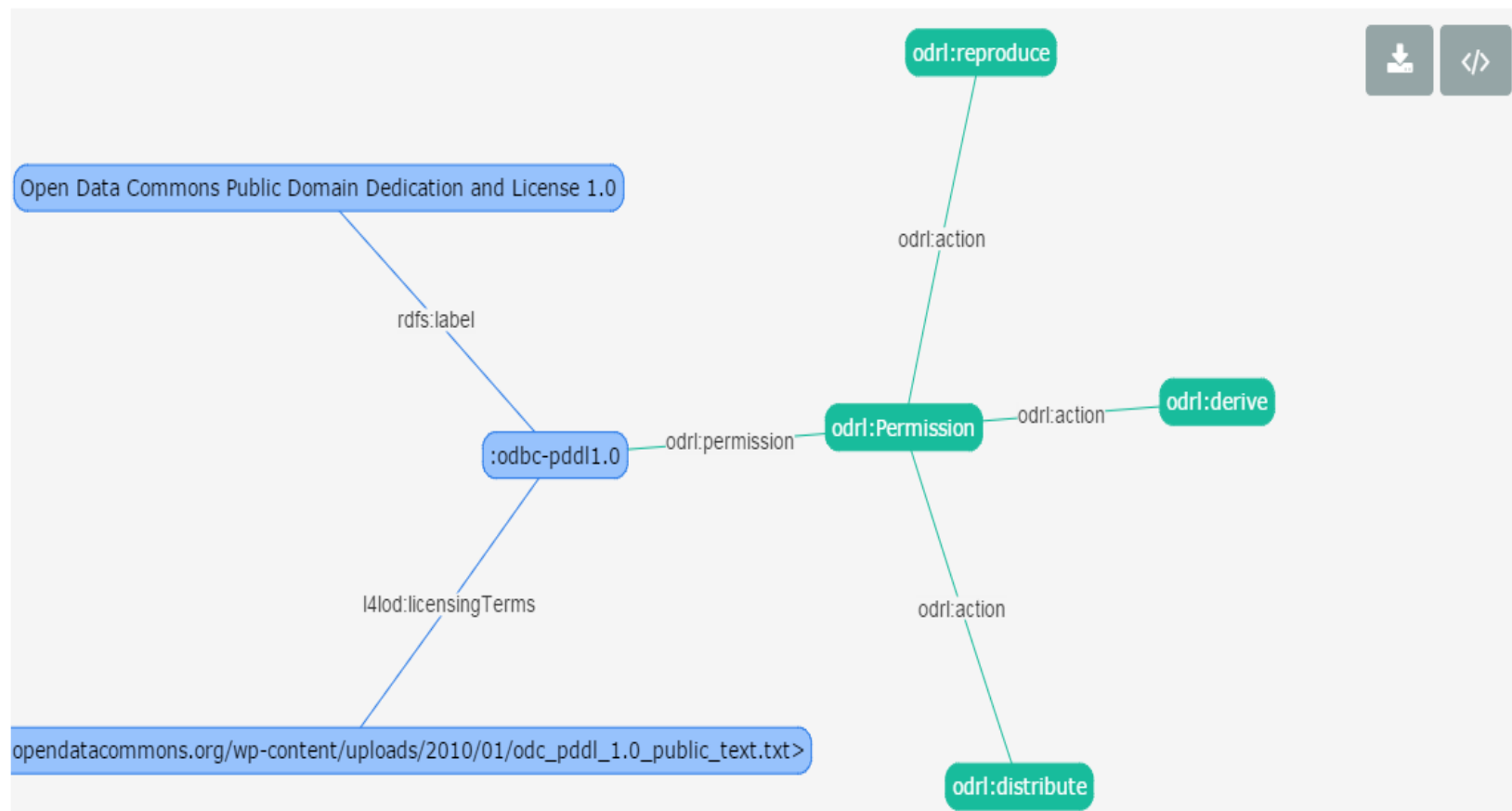
First element name (el_1)	Second element name (el_2)	Compatibility ($el_1 \circ el_2$) (\circ - compatibility operator, T - true, F - false)
Sharing	Derivative Works	T
Reproduction	Distribution	T
Reproduction	Derivative Works	T
Reproduction	Sharing	T
Distribution	Derivative Works	T
Sharing	Distribution	F

The **Licentia** tool



Licenses Visualizer

Open Data Commons Public Domain Dedication and License 1.0



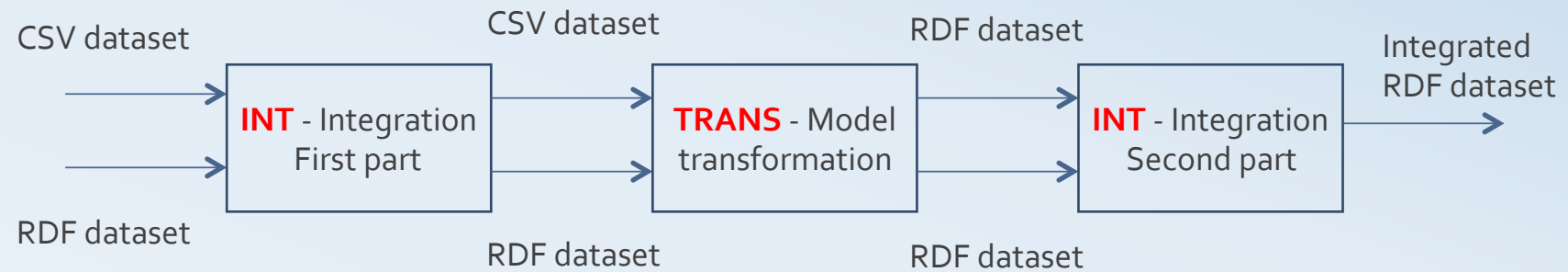


More on

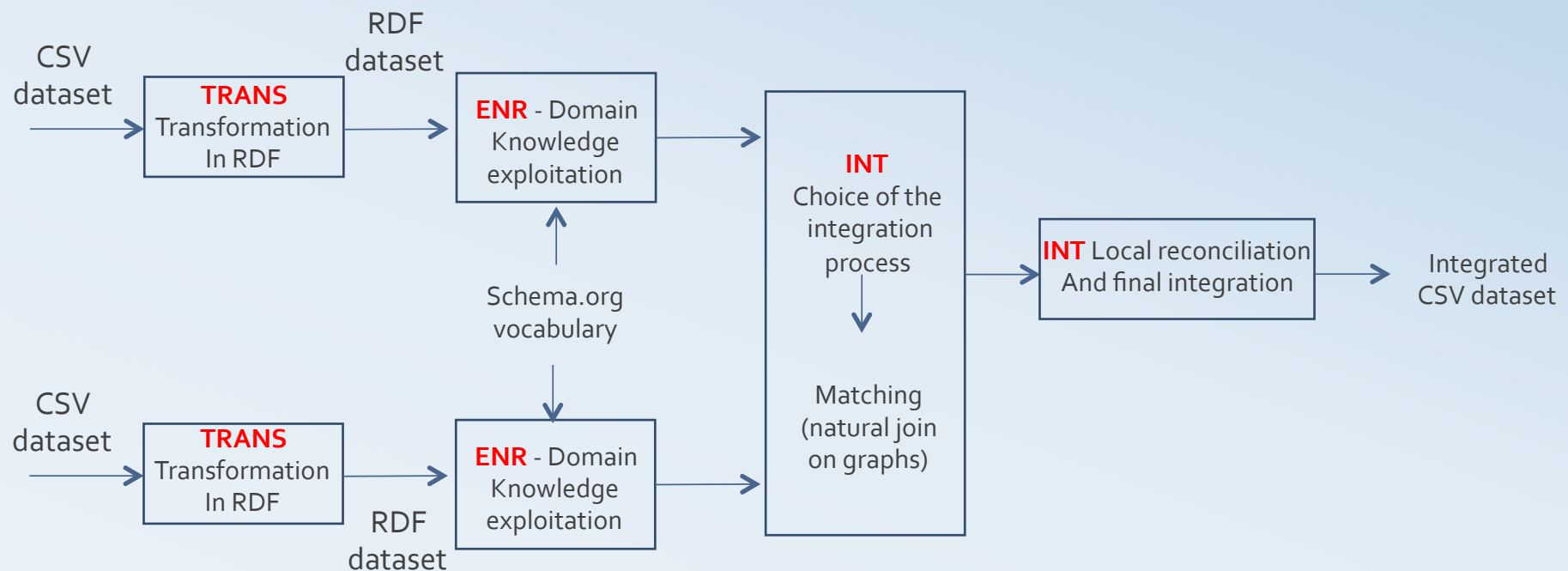
3. Choice of the integration process



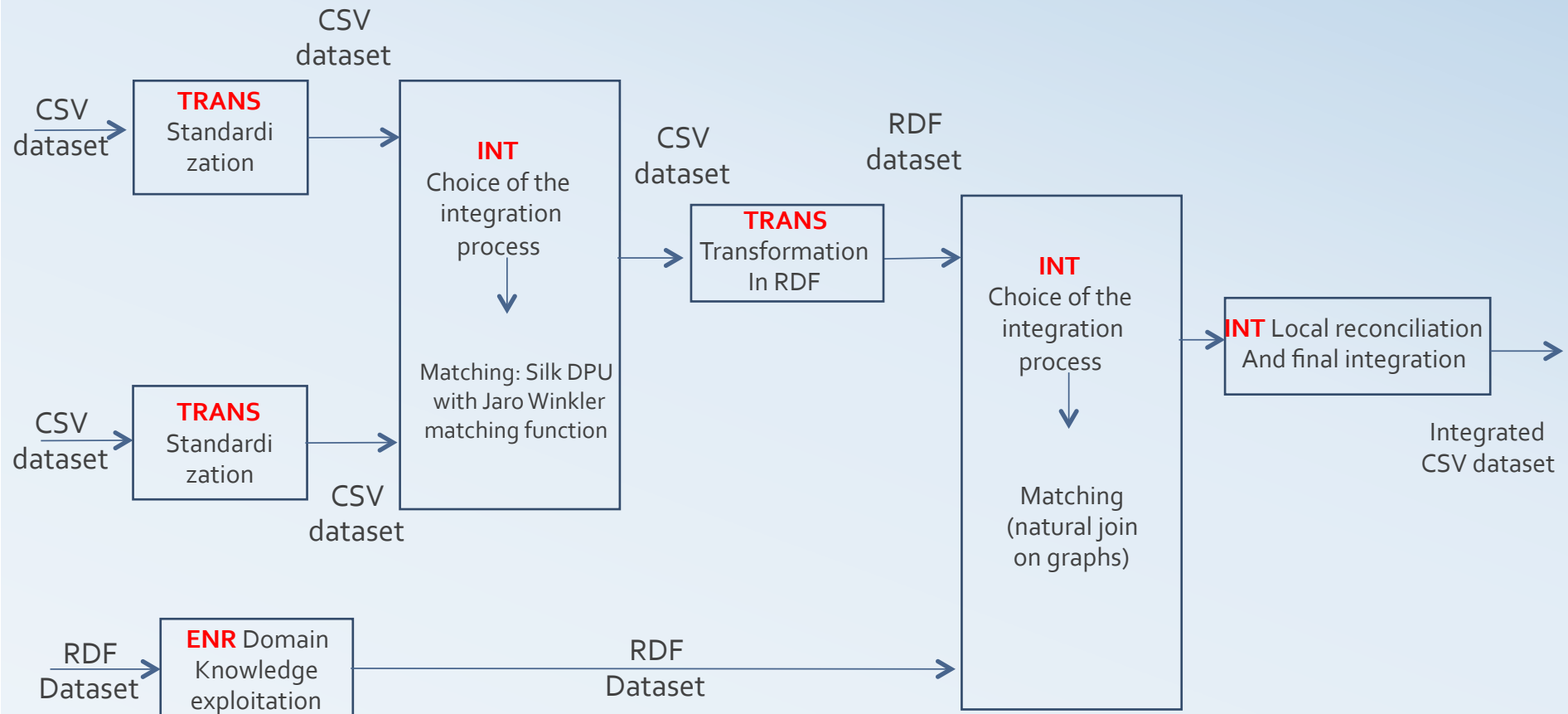
Example 1 of integration workflow: **CSV + RDF datasets**



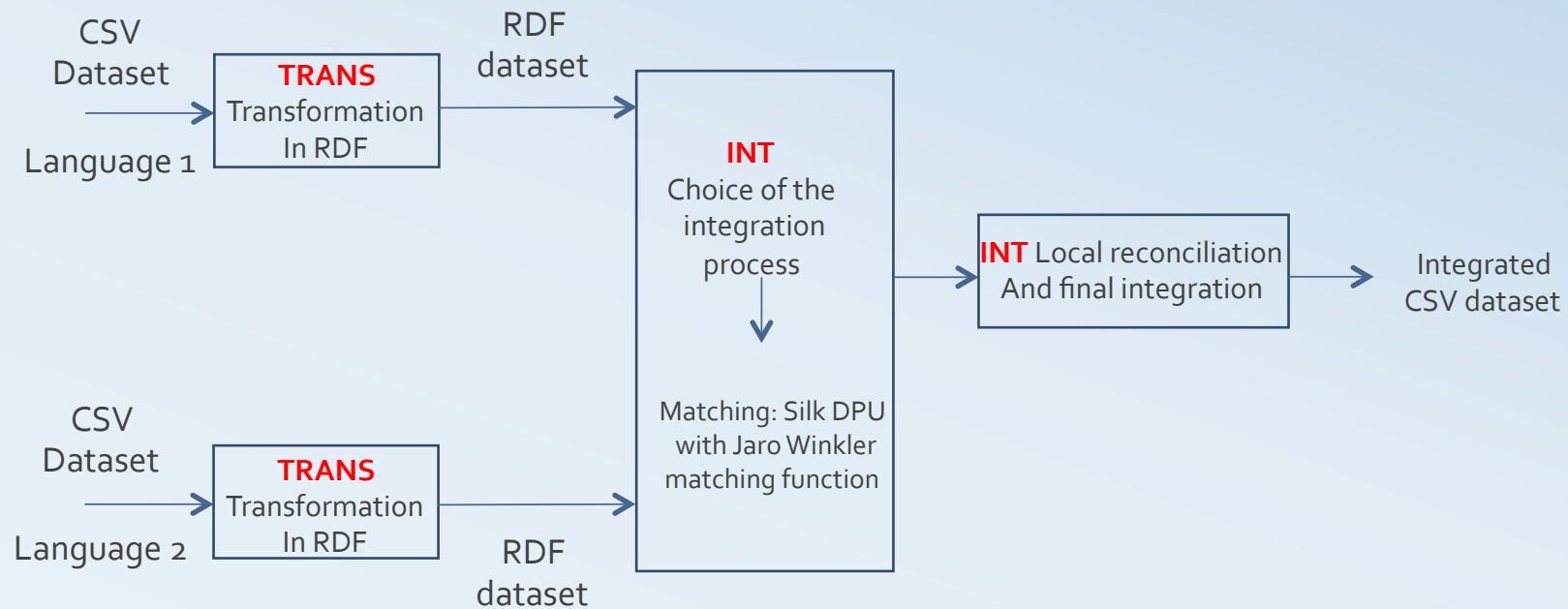
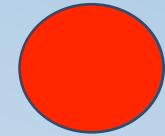
Example 2 of integration workflow: two CSV datasets adopting the same language (e.g. Italian)

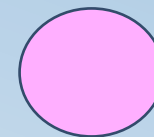


Example 3 of integration workflow: **two CSV datasets + one RDF dataset**



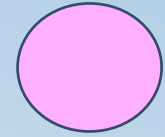
Example 4 of integration workflow: two CSV datasets adopting different but similar languages e.g. italian and spanish





More on 4. **Pre-integration**

4. Pre-integration activities



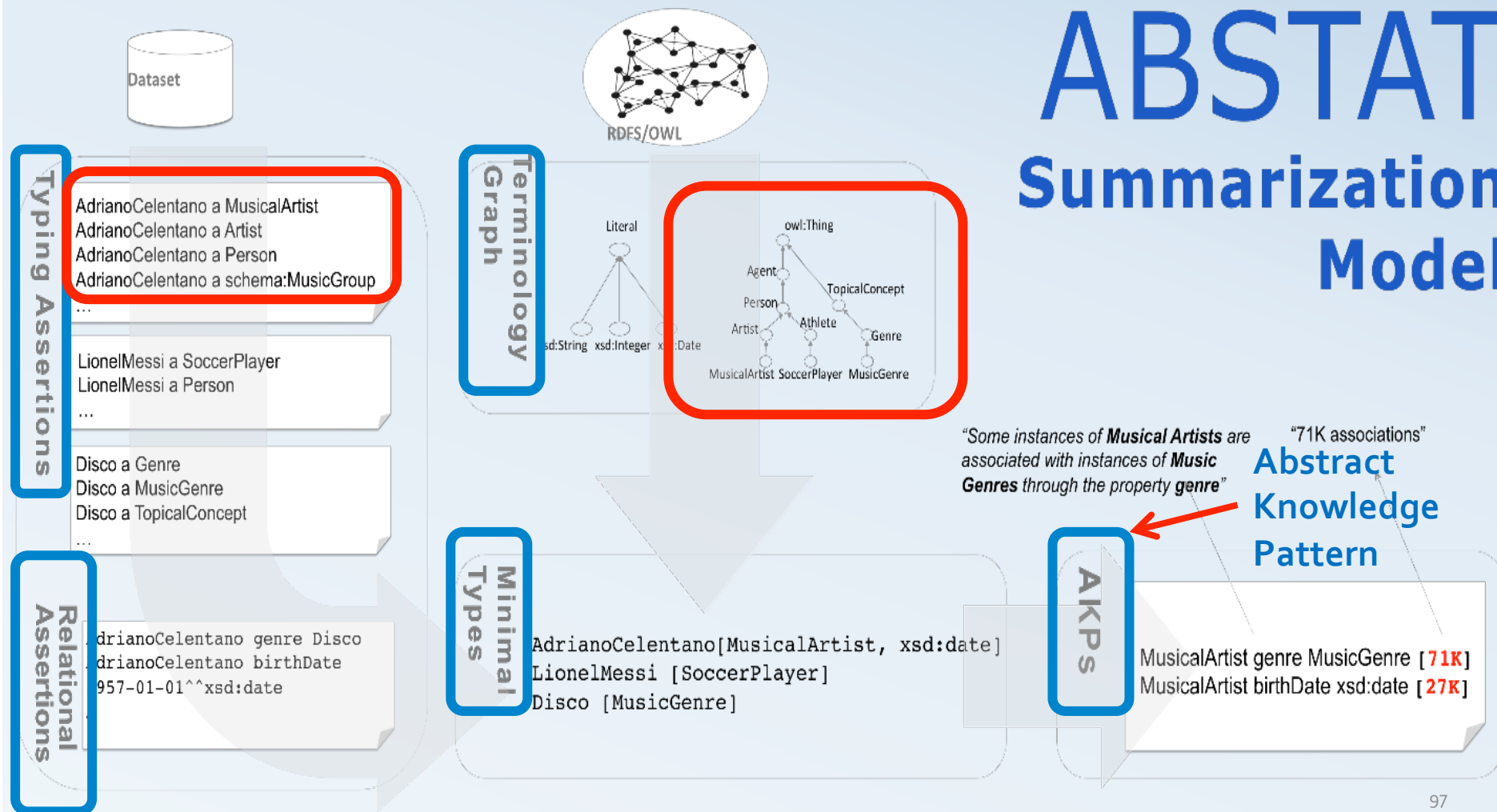
- **Data profiling**
- Assessment and improvement of quality (seen yet)
- **Normalization, standardization and restructuring** of datasets

4.1 Data profiling

ABSTAT, a linked data **summarization framework** for better understanding of a dataset

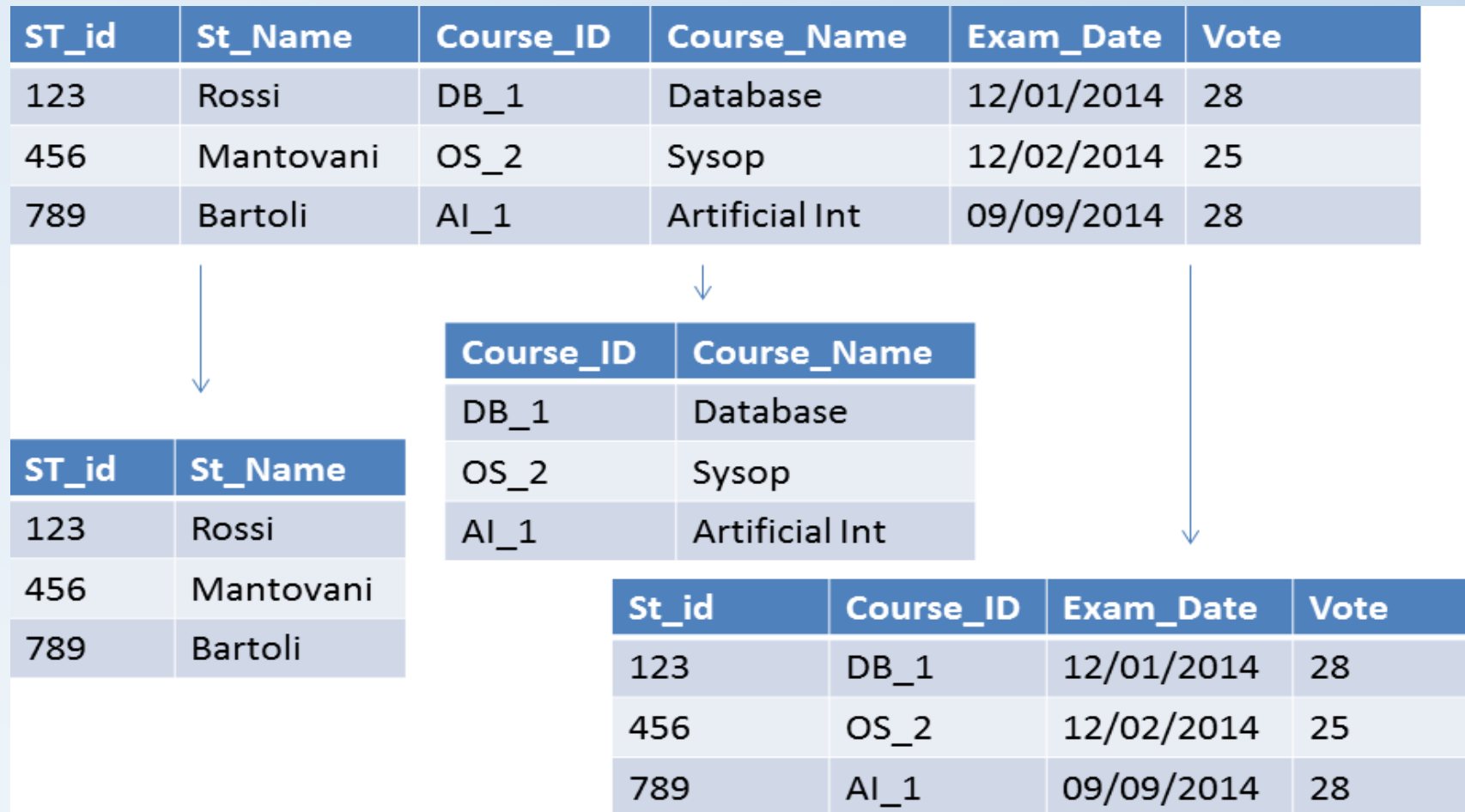
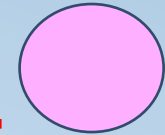


ABSTAT Summarization Model



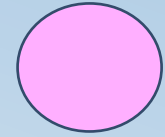
4.3 Normalization of datasets

Example of **transformation into BCNF**



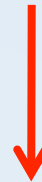
4.3 Restructuring of datasets

Example of **format reconciliation**



Dataset 1

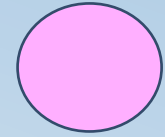
Location
550 First Avenue 10026 New York



Dataset 2

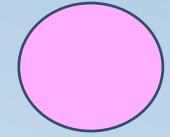
streetAddress	Number	ZipCode	City
First Avenue	550	10026	New York

6. Matching



- **At schema level** aims at finding **correspondences** between entities of two different schemas, in order to find equivalent, similar or, in general, corresponding entities.
- **At instance level**, it means to discover **matching and non-matching pairs**.

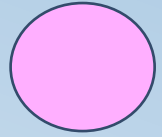
7. Local reconciliation and final integration



- Schema level **local reconciliation**
- Data level final integration (**fusion**)

7. Local reconciliation and final integration

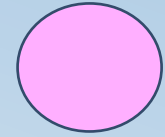
Fusion



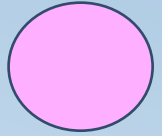
Strategy	Classification	Short Description
PASS IT ON	ignoring	escalates conflicts to user or application
C O N S I D E R A L L POSSIBILITIES	ignoring	creates all possible value combinations
TAKE THE INFORMATION	avoiding, instance based	prefers values over null values
NO GOSSIPING	avoiding, instance based	returns only consistent tuples
TRUST YOUR FRIENDS	avoiding, metadata based	takes the value of a preferred source
CRY WITH THE WOLVES	resolution, instance based, deciding	takes the most often occurring value
ROLL THE DICE	resolution, instance based, deciding	takes a random value
MEET IN THE MIDDLE	resolution, instance based, mediating	takes an average value
KEEP UP TO DATE	resolution, metadata based, deciding	takes the most recent value

From: Bleiholder, J., & Naumann, F. (2009). Data Fusion - ACM Comput. Surv., 41(1), 1:1–1:41. doi:10.1145/1456650.1456651.

8. Maintenance



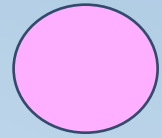
- When maintaining data, updates may create new heterogeneities among data sources, so that a new re-integration process should be taken into account.
- If we want to avoid the costly operation of running again the integration phases from scratch, two possible approaches are:
 - maintenance of data through **incremental integration**;
 - **rule-based maintenance** of data.



Let us go back to
3. Choice of integration process

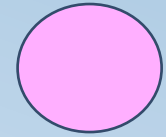
Let us go back to

3. Choice of integration process



Three cases: 1. **Local data set annotation**

1. Each dataset is annotated using a **local ontology**
2. These mappings are used to **extract RDF data** from each dataset.
3. The two **RDF datasets are mapped**, at the ontology and at the instance level, with the result that a set of correspondences between schema elements and entities of the two datasets are established.
4. These mappings are used to generate a (**virtual or materialized**) unique dataset that a user can query and interact with.



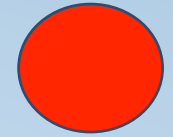
3. Choice of integration process

Three cases: 2. **Global dataset integration**

1. A new or existing **common ontology** is used in the dataset annotation process.
2. **Mappings from dataset cells to instances** described in the KB are found.
3. All the established **mappings are exploited** to extract **one new dataset** from the two.
4. The dataset output of the integration process **is described using the ontology** at the highest possible extent.

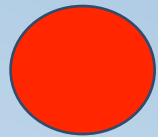
3. Choice of integration process

Three cases: 3. **GLocal dataset integration**



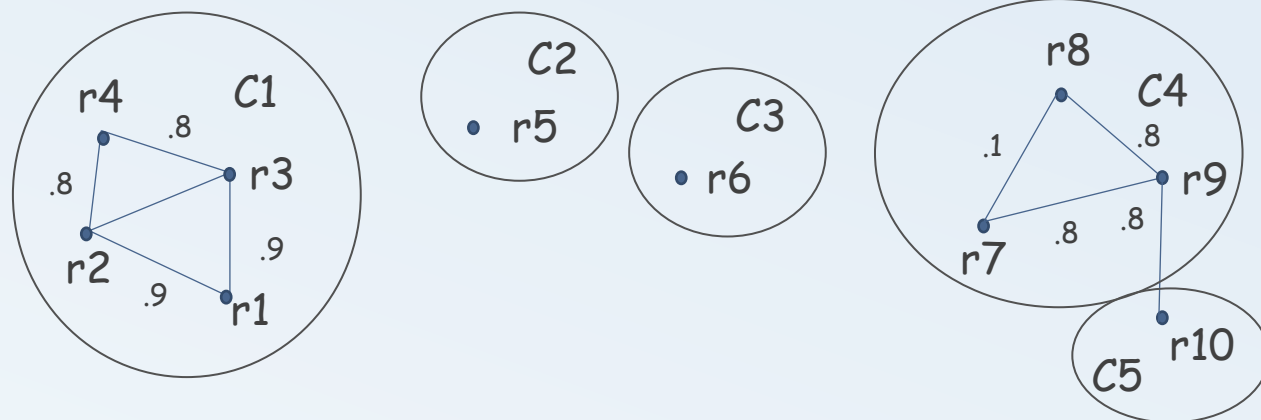
1. Only the **parts** of the input datasets that are **not covered by existing ontologies** are modeled.
2. The user is free to **use new user-defined ontology elements** to model his data when necessary or preferred.
3. The new ontology pieces that are designed in the integration tasks will be made available for future tasks.
4. **User-defined domain ontologies and mappings** from these ontologies to other ontologies used in web KBs **are created as different integration tasks are accomplished**.
5. The ontology and mapping creation process is *incremental and complementary* to the integration process.

8. Maintenance: **incremental integration** – example from C. Batini & M. Scannapieco - **Data and Information Quality**, Springer, forthcoming, November 2015

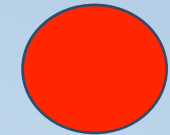


	BizId	Id	Name	Street address	City	Phone
D0	B1	r1	Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO	4155431510
	B1	r2	Starbucks	123 MISSION ST	SAN FRANCISCO	4155431510
	B1	r3	Starbucks	123 Mission St	SAN FRANCISCO	4155431510
	B2	r4	Starbucks Coffee	340 MISSION ST	SAN FRANCISCO	4155431510
	B3	r5	Starbucks Coffee	333 MARKET ST	SAN FRANCISCO	415534786
	B3	r6	Starbucks	MARKET ST	San Francisco	
	B4	r7	Starbucks Coffee	52 California St	San Francisco	4153988630
	B4	r8	Starbucks Coffee	52 CALIFORNIA ST	SAN FRANCISCO	4153988630
	B5	r9	Starbucks Coffee	295 California St	SAN FRANCISCO	415986234
	B5	r10	Starbucks	295 California ST	SF	

a. Original business listings

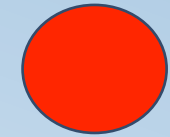


b. Matching results



8. Maintenance: **incremental integration** – example from C. Batini & M. Scannapieco - **Data and Information Quality**, Springer, forthcoming, November 2015

	BizId	Id	name	Street address	city	phone
D1	B6	r11	Starbucks Coffee	201 Spear Street	San Francisco	4159745077
D2	B3 B3	r12 r13	Starbucks Coffee Starbucks	MARKET STREET 333 MARKET ST	San Francisco San Francisco	4155434786 4155434786
D3	B1 B1	r14 r15	Starbucks Starbucks	123 MISSION ST STE ST1	SAN FRANCISCO San Francisco	4155431510 4155431510
D4	B5 B4	r16 r17	Starbucks Starbucks	295 CALIFORNIA ST 52 California St	SAN FRANCISCO SF	4155431510 4153988630



Rule based maintenance of data

Example From C. Batini & M. Scannapieco - **Data and Information Quality**, Springer, forthcoming, November 2015

Record	Name	Zip	Phone
r1	John	54321	123-4567
r2	John	54321	987-6543
r3	John	11111	987-6543
r4	Bob	null	121-1212

a. Records to match

Comparison Rule	Definition
B1	P_{name}
B2	$P_{name} \text{ AND } P_{zip}$
B3	$P_{name} \text{ AND } P_{phone}$

b. Evolving from rule B1 to rule B2

Contents



1. Introduction to Open Data and the COMSODE Publication Platform
2. Usage of COMSODE Techniques
3. The COMSODE Methodology: generalities
4. Methodology for one dataset
5. Methodology for multiple datasets
6. **Social value of open data**



Social value: from Economist October 2011

The Economist

World politics Business & finance Economics Science & technology Culture Blogs Debate & discuss Multimedia Print edition

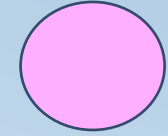
The Open Government Partnership
The parting of the red tape
Is it just another global talking-shop—or a fresh approach to shaking out government secrecy?

Oct 8th 2011 | NEW YORK AND TALLINN | from the print edition

Like 151 0

UGANDA is not best known as a testbed for new ideas in governance. But research there by Jakob Svensson at the University of Stockholm and colleagues suggested that giving people health-care performance data and helping them organise to submit complaints cut the death rate in under-fives by a third.

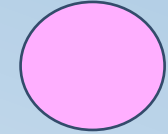
Publishing data on school budgets reduced the misuse of funds and increased enrolment.



What is **Social Value of Open Data**?

- Is the benefit that users of open data get with reference to their quality of life

How is **Social Value of Open Data** comparable and measurable?



We propose a methodology that compares and evaluates social value of open data sources, and we apply it to the specific domain of **hospital care**

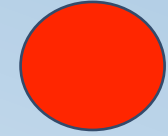
In our methodology, three dimensions of social value for open datasets are defined:

- Intensional social value
 - Extensional social value
 - Perceived social value
-
- In the following for brevity, we consider **intensional social value**



Example: Social Value of Health Data

- We want to **compare social value of health data** provided by the following sources
- US Hospitals
- Canada Hospitals
- Dove Salute IT – Ministry of Health, Italy
- On line Magazine Wired, Italy



Datasets Group Selection Example

Our sources

For US:

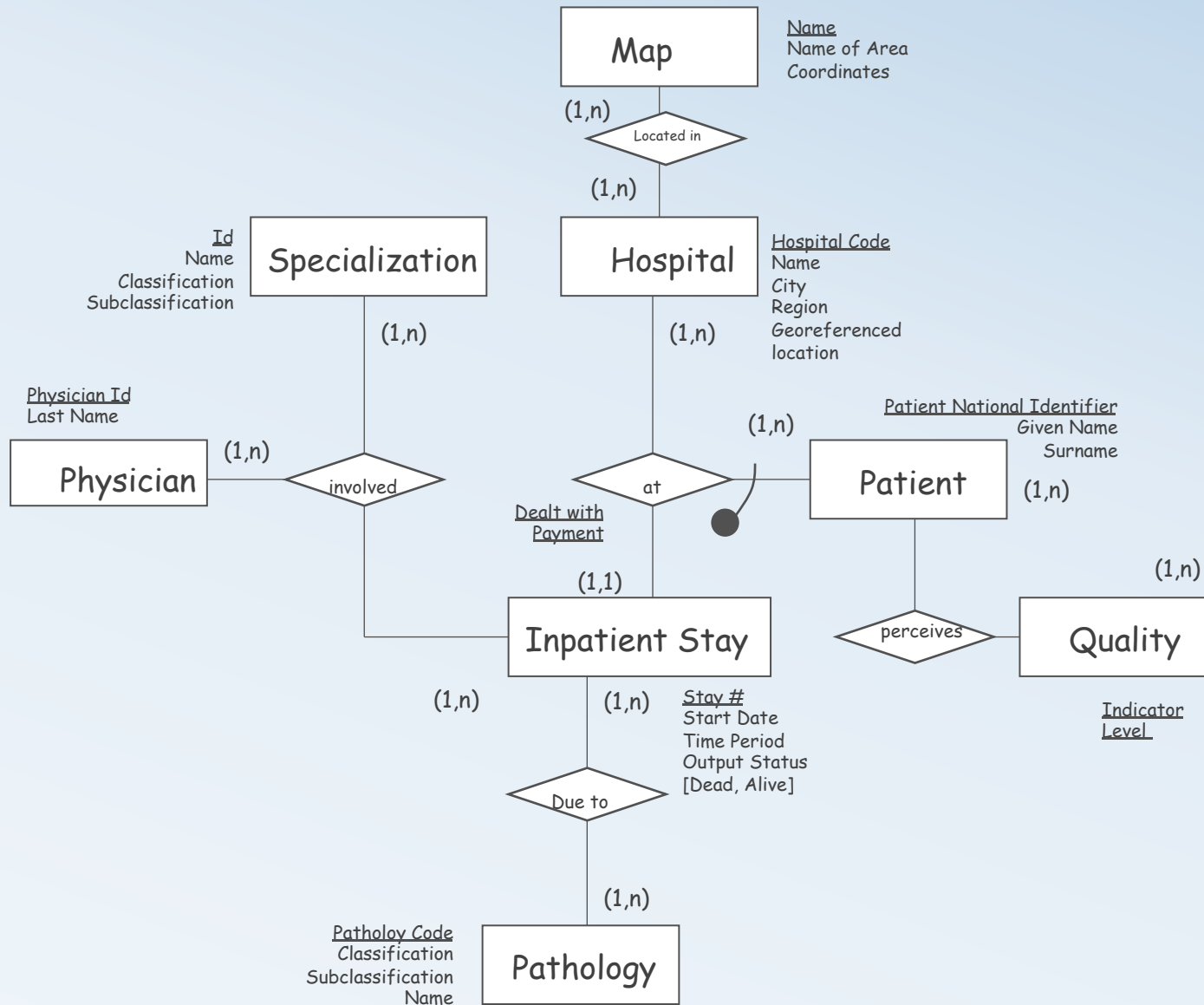
*data.cms.gov, health.data.ny.gov,
data.medicare.gov*

For Canada: *www.cihi.ca*

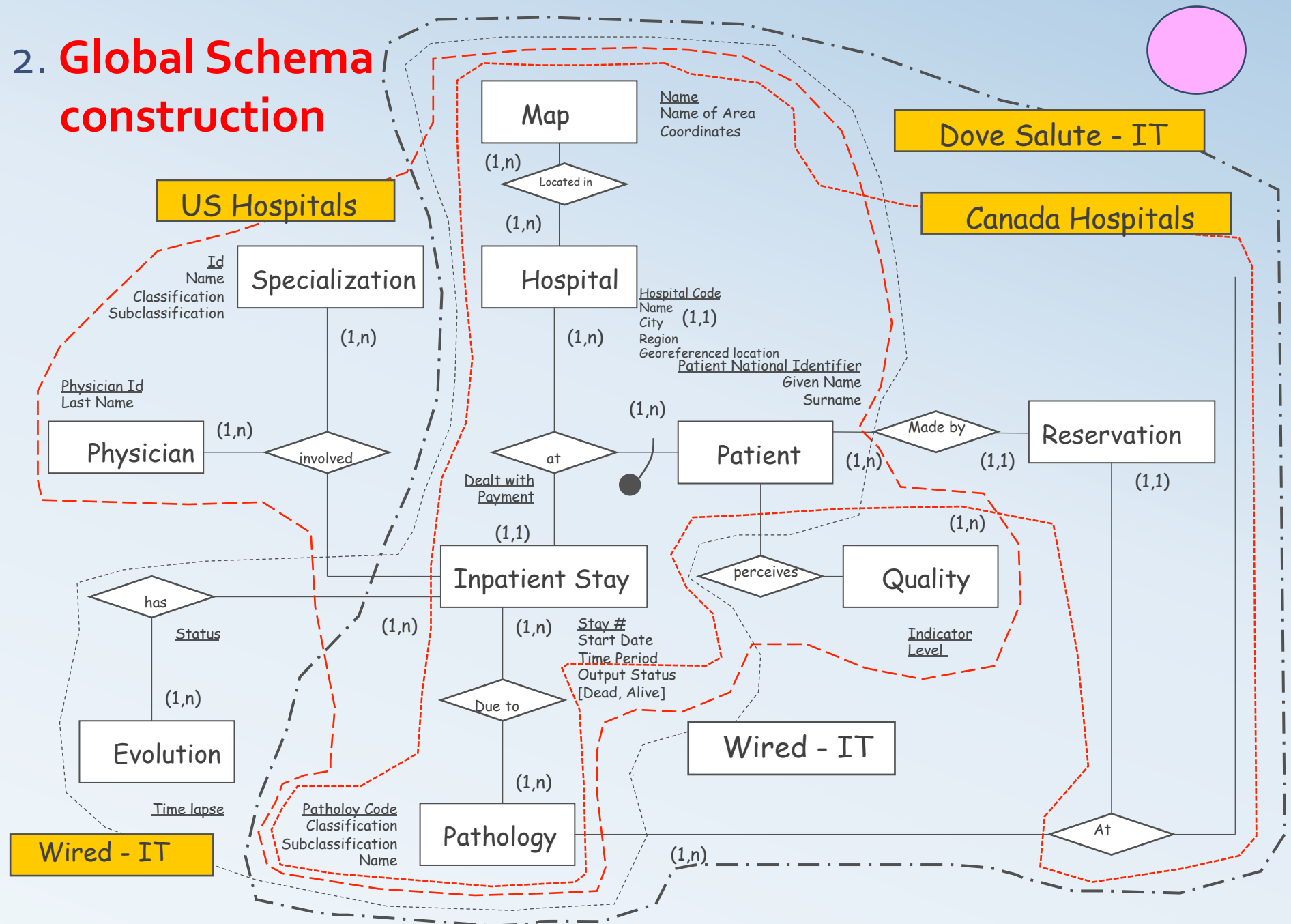
For Italy:

- *95.110.213.190/PNEed14,
dovesalute.gov.it*
- *www.wired.it*

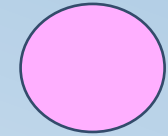
1. Single source **Integrated schema creation**: the case of US Hospitals schema



2. Global Schema construction

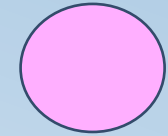


3. List of relevant events/queries



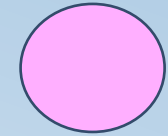
- Q1. Closest hospital
- Q2. Hospitals that take care of the disease
- Q3. Percentage of success of hospitals in taking care of the disease measured at discharge time
- Q4. Percentage of success of hospitals in taking care of the disease measured during the recovery
- Q5. Percentage of success of doctors in hospitals in taking care of the disease measured at discharge time
- Q6. Percentage of success of doctors in hospitals in taking care of the disease measured during the recovery
- Q7. For a given disease, reputation of hospitals perceived by hospitalized patients
- Q8. For a given disease, reputation of hospitals perceived by relatives of hospitalized patients
- Q9. Average wait time before admission by pathology

4. Mapping of Entities and Queries to Schemas



	Map	Hospital	Speciali zation	Patient	Physician	Reservation	Inpatient Stay	Quality	Pathology	Evolution
US	X	X	X	X	X		X	X	X	
Canada	X	X		X		X	X		X	
Wired - It	X	X		X			X		X	X
Dove Salute - IT	X	X		X		X	X	X	X	
Q1	o	o								
Q2		o							o	
Q3		o		o			o		o	
Q4		o		o			o		o	o
Q5		o	o	o	o		o		o	
Q6		o	o	o	o		o		o	o
Q7		o		o			o	o	o	
Q8		o		o			o	o	o	
Q9		o		o		o	o		o	

5. Evaluation of Coverage → Social Value



Data source	Score
US	7 on 9
Canada	6 on 9
Dove Salute	6 on 9
Wired	5 on 9



We have finished....

**COMSODE - Useful links
and resources**



COMSODE - Useful links

COMSODE EU Project

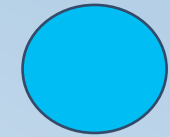
<https://www.comsode.eu/>

COMSODE news and main milestones

<https://www.comsode.eu/index.php/blog/>

COMSODE ODN official page

<http://opendatanode.org/product/open-data-node/>



COMSODE - Useful links

COMSODE ODN official DEMO page

<http://demo.comsode.eu/>

COMSODE ODN – Unified Views

<https://demo.comsode.eu/unifiedviews/>

COMSODE ODN Documentation

<https://utopia.sk/wiki/pages/viewpage.action?pageId=57049223>

UnifiedViews Documentation

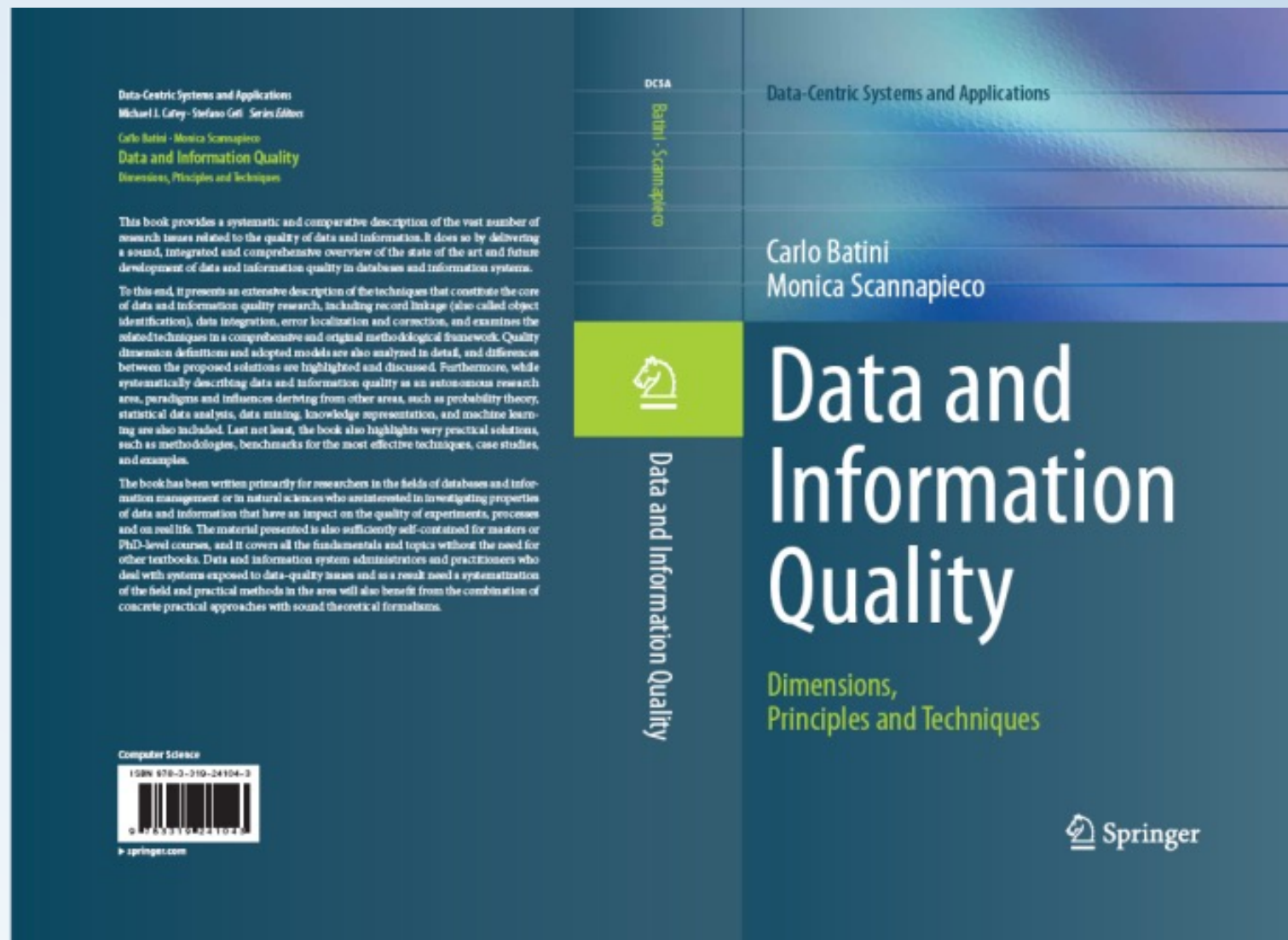
<https://grips.semantic-web.at/display/UDDOC/Introduction>



ABSTAT Useful Links

- ABSTAT Alpha
 - web app online (navigation + advanced search): abstat.disco.unimib.it
 - SPARQL endpoint: abstat.disco.unimib.it/sparql
 - full-text prototype: abstat.disco.unimib.it/search
- Description of ABSTAT summarization model
 - Matteo Palmonari, [Anisa Rula](#), [Riccardo Porrini](#), [Andrea Maurino](#), [Blerina Spahiu](#), [Vincenzo Ferme](#): ABSTAT: Linked Data Summaries with ABstraction and STATistics. [ESWC \(Satellite Events\) 2015](#): 128-132
- More datasets will be summarized soon
 - Ask for a dataset to summarize at palmonari@disco.unimib.it

Data and Information Quality forthcoming, November 2015



The End

Appendixes

Ao. Comparison of COMSODE with competition

Comparison - 1

Feature	Competition	ODN
Licencing	Usually provided as SAAS., some parts open source	Open source
Delivery model	Mostly SAAS, cloud bases	user installs ODN wherever and however appropriate, we plan also SaaS alternative
Availability of methodology	no	yes, solution accompanied with methodologies
support for RDF data publishing	varies, usually not focused on it	strong
Statistics of usage	yes	yes
Data sharing	yes	yes
Searching for data	yes	yes, but in fully free version (ODN only) only for metadata, full search available for combination of ODN+Search by Strategy
Visualization	yes	Yes – LDMI tool

Comparison - 2

Feature	Comepetition	ODN
API and other development stuff	yes	yes
Where is the data?	Usually on cloud (in their system	Wherever user chooses
Where will the money go?	Licenses - to the vendor, service fees - to the vendor or partner implementation services - to the vendor or partner, administration and operation - to the vendor or partner	Licenses - for free, no service fees, implementation services - to the vendor or partner, users may implement themselves, administration and operation - to the vendor or partner, users may do it themselves
Vendor lock-in	Typically high, caused mostly SaaS nature of solutions. User is dependant on the service provider and doesn't have control over stored data.	Low, as it is typical for open source software. User has a lot of options how to implement and run ODN.

A1. COMSODE Requirements

Open Government Data publication methodology requirements

RQ2 - Assessment of demand for OGD:

- ❖ In order to focus the effort on datasets that are in demand by the potential users, the OGD publication methodology should provide guidelines how to assess the demand for OGD.

Open Government Data publication methodology requirements

RQ₃ - Selection and prioritization of datasets:

- ❖ Public sector bodies own various datasets. Therefore the OGD publication methodology should provide guidelines how to identify and select suitable datasets for opening up.
- ❖ Because it might not be always feasible to publish all the selected datasets at once, recommendations on prioritization of datasets for publication should be provide as well.

Open Government Data publication methodology requirements

RQ4 - OGD benefits assessment:

- ❖ Because the OGD benefits are not always systematically assessed, the OGD publication methodology should provide explanation of the typical OGD benefits and recommendations for their assessment

Open Government Data publication methodology requirements

RQ5 - Effort and costs estimation:

- ❖ In order to support financial and resource management of the OGD publication, recommendations on effort and cost estimation should be provided by the OGD publication methodology

Open Government Data publication methodology requirements

RQ6 - Recommendation about fees:

- ❖ In order to align the OGD publication with the applicable charging principles the OGD publication methodology should discuss the issues of collecting fees for data vs. providing data for free.

Open Government Data publication methodology requirements

RQ7 - Ensuring compliance with the legislation:

- ❖ The OGD publication methodology should provide guidelines or recommendations about analysis of the applicable legislation and the possible limitations to publication of particular datasets.
- ❖ The goal of this analysis is to ensure that publication of the selected datasets complies with the applicable legislation. In case that it is not possible to publish the primary data (e.g. due to the personal data protection), the methodology should provide recommendations about anonymization of the data.

Open Government Data publication methodology requirements

RQ8 - Risk analysis:

- ❖ In order to properly manage the possible risks associated with the OGD publication, the methodology should discuss the risk management issues and it should provide recommendations how to deal with common risks .

Open Government Data publication methodology requirements

RQ9 - Licensing:

- ❖ In order to ensure that the published data are legally open, the methodology should provide recommendation about licensing Open Data and how to inform users about the terms of use of the datasets

Open Government Data publication methodology requirements

RQ10 - Reuse of already published datasets:

- ❖ In order to avoid unnecessary duplicate publication of datasets on different web sites, the methodology should provide recommendations how the OGD publishers should reuse existing datasets and inform them about existence of the data portals

Open Government Data publication methodology requirements

RQ11 - Recommended data formats:

- ❖ In order to ensure that the published data is technically open, the methodology should provide a set of recommended data formats together with guidelines for their application

Open Government Data publication methodology requirements

RQ12 - Interlinking of related datasets:

- ❖ Datasets might relate to each other. In order to increase value of the provided datasets, the methodology should provide recommendations or guidelines how to interlink datasets and how and when to publish Linked Open Data

Open Government Data publication methodology requirements

RQ13 - ICT impact assessment:

- ❖ Publication of OGD might require changes to the underlying ICT infrastructure and systems. Therefore ICT impact assessment should be taken into consideration by the OGD publication methodology

Open Government Data publication methodology requirements

RQ14 - OGD publication process:

The OGD publication process consist of the four phases:

- ❖ Development of open data publication plan,
- ❖ Preparation of publication,
- ❖ Realization of publication and
- ❖ Archiving.

Each of the phases is made of set of tasks, under the responsibility of specific roles.

Open Government Data publication methodology requirements

RQ15 - Data cataloguing:

- ❖ In order to ensure that the published data are easily located by their potential users and that the information needed to understand the data schema, semantics and possible limitations are available to the users, guidelines on data cataloguing should be provided by the methodology together with the recommended metadata schema.
- ❖ Metadata quality assurance should be discussed as well.

Open Government Data publication methodology requirements

RQ16 - Data quality assurance:

- ❖ In order to ensure that the published open datasets meet the quality requirements, the methodology should provide guidelines for data quality assurance

Open Government Data publication methodology requirements

RQ17 - Ensuring easy access to datasets:

- ❖ The methodology should discuss the common barriers to access to the datasets including a possible pitfall of required user registration at data portals. This should ensure that the access to the published datasets is as easy as possible.

Open Government Data publication methodology requirements

RQ18 - Dataset maintenance:

- ❖ Dataset maintenance should be addressed by the OGD publication methodology in order to ensure that the datasets are regularly updated

Open Government Data publication methodology requirements

RQ19 - Communication strategy:

- ❖ Recommendation how to develop OGD communication strategy should be provided by the methodology.
- ❖ This strategy should define how the feedback from the users will be processed, how the OGD will be promoted and how the OGD activities of the publisher will be aligned with other relevant OGD initiatives.

Open Government Data publication methodology requirements

RQ20 - Independence on the central data portal

- ❖ Even though many countries have launched their national OGD portal there might be other countries where the central data portal is not available. Therefore the recommendations and practices described in the OGD publication methodology should be independent on the central (national) data portal.

Open Government Data publication methodology requirements

RQ21 - Recommended software:

- ❖ In order to ease selection of the suitable software tools for publication of OGD, the methodology could provide a list of recommended software tools.

Open Government Data publication methodology requirements

RQ22 - Public sector bodies of different size should be taken into consideration:

- ❖ Public sector consists of different public sector bodies. The small ones might have a very limited resources for publication of OGD. Therefore size of the public sector bodies should be taken into consideration when formulating recommendations of the OGD publication methodology.

A2. COMSODE Phases and tasks

Tasks for Phases open data publication plan

1. Analysis of data sources
2. Identification of datasets for opening up
3. Determination of target level of openness
4. Effort estimation
5. Definition of the open data publication plan

Tasks for Phases

Preparation for publication

1. Data sources access configuration
2. Definition of the catalogue record schema and the target data catalogues
3. Description of the datasets
4. Selection and implementation of the software tools
5. Definition of the open data publication plan
6. Definition of the approach to the dataset publication
7. Design and implementation of the ETL procedures
8. Testing of the ETL procedures
9. Licensing

Tasks for Phases Realization of publication

1. Initial publication of the dataset
2. Data cataloguing
3. Dataset maintenance

Tasks for Phases Archiving

1. Termination of the dataset maintenance
2. Termination of the dataset publication

A₃. COMSODE Artifacts

Artifacts of the Methodology

Po1 - Development of the Publication Plan (18 artifacts in total)

Po2 - Phase: Preparation of Publication (21 artifacts in total)

Po3 – Realization of Publication (6 artifacts in total)

Po4 – Phase Archiving (6 artifacts in total)

Artifacts of the Methodology

CCA₁: Data Quality Management (13 artifacts in total)

CCA₂: Communication Management (19 artifacts in total)

CCA₃: Risk Management (6 artifacts in total)

CCA₄: Benefits Management (6 artifacts in total)

Artifacts of the Methodology

Information strategy (Information policy) of the organization

Type: I – Po1

Artifacts of the Methodology

Organizational structure

Agendas (public services) of the organization

Type: I – Po1 –CCA2

Artifacts of the Methodology

List of information systems and databases of the organization

Type: I – P01 – P02

Artifacts of the Methodology

Internal directives

Type: I – P01 – P02 – CCA2

Artifacts of the Methodology

FOI requests

Type: I – Po1 – CCA2

Artifacts of the Methodology

Relevant legislation and strategic documents

Type: I – Po1 – Po2 – CCA1 – CCA4

Artifacts of the Methodology

Map of data sources and datasets

Type: I – Po1 – Po2 – CCA4

Type: O – Po1

Artifacts of the Methodology

List of datasets for opening up

Type: I – P₀₁ – P₀₂ – CCA₁ – CCA₂ – CCA₃ – CCA₄

Type: O – P₀₁

Artifacts of the Methodology

Effort analysis documentation

Type: I – Po1

Type: O – Po1

Artifacts of the Methodology

Open data publication plan, structure:

- ❖ goals of the organization which will be achieved by opening up the data
- ❖ map of data sources and datasets
- ❖ the list of datasets, including the identified benefits and risks, levels of openness and estimated effort
- ❖ priorities
- ❖ publication schedule
- ❖ assignment of workers to roles

Type: I – Po2 – CCA2

Type: O – Po1

Artifacts of the Methodology

Data sources accesibility report:

Type: I – P01 – P02

Type: O – P02

Artifacts of the Methodology

Catalogue record schema

Target data catalogue

Type: I – P₀₂

Type: O – P₀₂

Catalogue records

Type: I – P₀₂ – P₀₃ - P₀₄

Type: O – P₀₂ – P₀₄

Artifacts of the Methodology

Dataset schema

Type: I – P₀₁ – P₀₂ – CCA₁

Type: O – P₀₁ – P₀₂

Artifacts of the Methodology

Ontology

Type: I – P₀₂ – CCA₁

Type: O – P₀₂

Artifacts of the Methodology

Architecture of the software infrastructure for open data publication

Software infrastructure implementation report

Type: O – Po2

Artifacts of the Methodology

ETL design documentation

Type: I – P02

Type: O – P02

ETL procedures and the related documentation

Type: I – P02 – P03 – CCA1

Type: O – P02

Artifacts of the Methodology

ETL procedures test documentation

Type: I – P02 – P03 – CCA1

Type: O – P02

ETL procedures test results

Type: O – P02

Artifacts of the Methodology

Terms of use / licence

Type: I – P03 – CCA1

Type: O – P02

Artifacts of the Methodology

Published catalogue records

Type: I – P03

Type: O – P03

Artifacts of the Methodology

Published datasets

Type: I – P₀₄

Type: O – P₀₃

Artifacts of the Methodology

Test data

Type: I – P_{O2}

Type: O – P_{O2}

Artifacts of the Methodology

Accesible primary (source) data

Type: I – P02

Type: O – P02

Artifacts of the Methodology

Dataset maintenance termination report

Unmaintained datasets

Dataset publication termination report

No longer published datasets

Type: I – CCA₂

Type: O – PO₄

Artifacts of the Methodology

Potential user groups

Type: I – CCA₁ - CCA₂ – CCA₄

Type: O – CCA₂

Artifacts of the Methodology

Communication strategy

Type: I – P01 – CCA2

Type: O – CCA2

Communication campaign supporting materials

Ready communication channel

Type: I – CCA2

Type: O – CCA2

Open data communication report

Type: O – CCA2

Artifacts of the Methodology

Disseminated information about open data

Type: O – CCA₂

Artifacts of the Methodology

User feedback

Type: I – P₀₁ – P₀₂ – P₀₃ – CCA₁ – CCA₂

Type: O – CCA₂

Artifacts of the Methodology

Documentation of the user feedback and data demand assessment

Type: I – Po1 – CCA₁

Type: O – CCA₂

Artifacts of the Methodology

Risk mitigation plan

Type: I – Po1 - CCA3

Type: O – CCA3

Incident report

Risk mitigation action implementation report

Risk register

Type: I – CCA3

Type: O – CCA3

Risk status report

Type: O – CCA3

Artifacts of the Methodology

Benefits management plan

Type: I – Po1 - CCA₄

Type: O – CCA₄

Benefits register

Type: I – CCA₄

Type: O – CCA₄

Open data benefits report

Type: O – CCA₄

Artifacts of the Methodology

Annual report

Type: I – Po1

Artifacts of the Methodology

Data sources quality requirement

Type: I – CCA₁

Type: O – CCA₁

Data sources quality assessment

Documentation of causes of errors

Processes for improvement

Type: O – CCA₁

A4. Roles of the Methodology

Roles of the Methodology

Owner

- ❖ A person or an entity that owns a dataset, i.e. s/he holds rights to the dataset or s/he is legitimate to make decisions about the dataset.
- ❖ Owner is legitimate to decide that a certain dataset will be published as open data and s/he is also legitimate to licence the dataset.

Roles of the Methodology

Curator

- ❖ A person or an entity that curates or maintains a dataset and its catalogue record (metadata). Curator keeps the published datasets accurate and up-to-date. Curator is usually appointed to his or her role by the data owner.

Roles of the Methodology

OD Catalogue Owner

- ❖ OD Catalogue Owner is an organization responsible for the Open Data catalogue. It is responsible for defining policies and rules governing cataloguing and its use. OD Catalogue Owner also selects the data cataloguing software.

Roles of the Methodology

OD Coordinator

- ❖ OD Coordinator is a person in an organization that is responsible for coordination and management of the open data related activities of the organization.

Roles of the Methodology

OD Catalogue Publisher

- ❖ OD Catalogue Publisher is a person or an organization that makes the data catalogue available to the potential users. OD Catalogue Publisher is responsible for operating the data catalog and s/he ensures maintenance of the underlying IT infrastructure.

Roles of the Methodology

IT Professional

- ❖ IT Professional is a person with skills and knowledge in the domain of the information and communication technologies.
- ❖ S/he provides support to other roles, develops and tests the ETL procedures and performs the transformation of the data into the target data formats.
- ❖ If Linked Open Data is published, IT Professional should have the necessary skills and knowledge for publication of LOD.

Roles of the Methodology

Data Quality Manager

- ❖ Data Quality Manager is a person with skills and knowledge in the domain of data quality.
- ❖ S/he is in charge of supervising all data quality components and the data quality lifecycle.
- ❖ Typically s/he is one of the supporters of the OD coordinator.

Roles of the Methodology

Data Quality Expert

- ❖ Data Quality Expert is a person with skills and knowledge in the phases and tools of the data quality life cycle.
- ❖ S/he is in charge of analyze, apply and in case create the ETL components related to the data quality.
- ❖ In case of small organization such role can be performed by the data quality manager itself.

Roles of the Methodology

Legal Expert

- ❖ Legal expert is a person with skills and knowledge in the domain of law and legislation.
- ❖ S/he provides his or her expertise about licencing of open datasets and s/he is involved in analysis of datasets where his or her expertise is required in order to ensure that the publication of datasets comply with the legislation.

A5. Tasks

Phase: Preparation for publication

Data quality requirements analysis

Phase: Preparation for publication

Quality assessment

Quality improvement

Tasks for Cross-cutting Activities

Data quality management

Phase: Preparation for publication

Data quality requirements analysis

Phase: Preparation for publication

Quality assessment

Quality improvement

Tasks for Cross-cutting Activities

Communication management

Identification of the potential user groups

Definition of the communication strategy

Engaging users during development of the OD publication plan

Phases involved (Development of the Open Data Publication Plan)

Setting up the communication channels defined in the communication strategy

Phases involved (Preparation for Publication)

Tasks for Cross-cutting Activities

Communication management

Preparation of the communication campaign

Informing about progress

Phases involved (Preparation for Publication)

Informing about open data

Analysis of the user feedback

Informing about the termination of maintenance or publication of a dataset

Phases involved (Realization of Publication)

Tasks for Cross-cutting Activities

Risk management

Identification and analysis of the potential risks

Definition of the risk mitigation plan

Phases involved (Development of the Open Data Publication Plan)

Update of the risk register with the information acquired during the preparation of the datasets

Realization of the risk mitigation actions relevant to the preparation of the datasets

Phases involved (Preparation of Publication)

Tasks for Cross-cutting Activities

Risk management

Risk management

Responding to events

Reporting about the state of the open data related risks

Phases involved (Realization of Publication)

Risk management

Responding to events

Reporting about the state of the open data related risks

Phases involved (Archiving)

Tasks for Cross-cutting Activities

Benefits management

Identification and analysis of the potential benefits

Definition of the benefits management plan

Phases involved (Development of the Open Data Publication Plan)

Update of the benefits management plan according to the information acquired during the preparation of the datasets

Phases involved (Preparation for Publication)

Tasks for Cross-cutting Activities

Benefits management

Benefits monitoring and management

Reporting about the open data related benefits

Phases involved (Realization of Publication)

Benefits monitoring and management

Reporting about the open data related benefits

Phases involved (Archiving)