METRIC FOR **DA**TA REUSABILITY (**MELODA**). 16 October, 2016

Version 4.10

1.- INTRODUCTION. BACKGROUND

1.1.- Mission. What MELODA is for ...

MELODA (**ME**tric for re**L**easing **O**pen **DA**ta) helps data publishers (public and private) to make the most of the data they release for reuse but also for data reusers to know what will be more usable for creating new products and services.

The goal is to make reuse as widely adopted as possible, including commercial uses of data. Every type of generated data is possible to be assessed wit **MELODA**, including open government data, scientific underlying data, crowdsourced data, etc.

So that legal issues about information are considered as well as technical standards, access issues, timeliness, geolocation and data model issues. With this goal in mind **MELODA** helps to those entities trying to impact on its surrounding societies to reach the most of the reuse of their information.

Although it is mostly useful for open data, if the data are not open it could help as well, however reusability of closed data could involve additional dimensions not considered here.

1.2.- What MELODA is not for…

MELODA by itself does not qualify full opendata portals, but unique datasets[1]. Some academic studies, though are using **MELODA** average on datasets to account for open data portals reusability.

1.3.- Users

This metric is aimed at those responsible people of repositories of information either open data portals, smart cities, scientific repositories, etc. As long as they release data and they want to maximize the use of data published under an open re-use scheme.

1.4.- Where to find

We release new versions of this metric in the official site http://meloda.org and some comments at my blog at http://gobernamos.com check it out.

1.5.- Check it out

There is a form in this link to check an actual dataset and get the results.

---

[1]See proposed definitions in next section

2.- DEFINITIONS

Dataset

Dataset: Group of structured data retrievable in a single instruction or link as a whole, with updating frequency larger than a once a minute.

DataJet

Group of structured data retrievable in a single instruction or link as a whole, with updating frequency equal or shorter than a once a minute.

3.-DIMENSIONS OF ANALYSIS

MELODA is focus on helping their users. It must obtain a figure by assessing the available information (metadata) of a dataset. This metadata has to be public through the web without (mostly) any further requirements. **MELODA** strictly refuses to qualify any dataset based on some features of the reuser.

In this version **MELODA** analyses six dimensions:

- Legal framework
- Technical Standards
- Accessibility to the information
- Data model sharing
- Geolocated information
- Real-time information

For each dimension up to 5 levels or reusability are considered.

Depending on the dimension different weights are assigned for every level.

2.1.- Legal dimension

This dimension considers five levels. Other legal considerations (i.e. Restriction of malicious use of information, misuse, requirements to include the last updated date, etc are not included).

Attribution to the original source is always considered as a must.

Description	Weight (%)
Level 1.Copyright. Data sources in this level: Restrict reuse without written agreement. It restricts unauthorized use Examples: 1) Copyright	0
Level 2. Private use Data sources in this level will allow the use of the data without approval processes but only for private uses. Not public use available. Examples:	10
Level 3. Non-commercial reuse Data sources in this level will allow reuse of data but they do not allow commercial uses of the information re-use. Examples: 1) (CC BY-NC-ND 4.0) Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 2) (CC BY-NC 4.0) Attribution-NonCommercial 4.0 International 3) (CC BY-NC-SA 4.0) Attribution-NonCommercial-ShareAlike 4.0 International	25
Level 4. Commercial reuse Data sources in this level will allow re-use of data, including the commercial reuse. (i.e. For any company in order to create new visualization of data) Examples: 1) (CC BY-ND 4.0) Attribution-NoDerivatives 4.0 International 2)Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)	90
Level 5. No restrictions or only attribution Data sources in this level will only ask re-users the attribution of the data. Examples: 1) Attribution 4.0 International (CC BY 4.0)	100

2.2.- Technical Standards

Description	Weight (%)
Level 1. Closed standard non reusable Data sources in this level are release on proprietary standards not focused on data reuse (i.e. image format for data or document standards). Examples: 1) pdf image 2) doc	10
Level 2. Closed standard reusable. Other closed formats somehow suitable for re-utilisation (data proprietary formats 3) shp 4) xls y xlsx with macro/formula	35
Level 3. Open standard Data sources in this level are published on open standards but as individual files. Examples: 1) csv without further information 2) txt 3) odb 4) odt 5) ods 6) WMS 4) xls y xlsx without macro/formula	60
Level 4. Open standard, individual metadata Data sources in this level includes those which provides metadata attached to any data. Semantic data meet this level Examples: 1) rdf, 2) rss, 3) json 4) xml (with meta-tags)	100

2.3.- Access to information

In any case the free (no cost) access is a must for being considered in this dimension. Otherwise punctuation would be 0.

Description	Weight (%)
Level 1. No web access or manual request Access to information requires a non-automatic approval process for access to data information or to register data in a non digital form. Examples 1) There is needed a non automatic mail or a in person interaction to get access	0
Level 2. Web Access URL with registration or with web interaction Access to information via the web, but requires user interaction to select the data source. Examples 1) It is needed to manually interact to get access 2) There is a limitation in the # of data to be accessed in one shot	10
Level 3. Web access or unique URL parameters to dataset Access to information via the web, but allows each of the datasets to be accessed individually, or through a unique URL, or by specific parameters in the query call Examples 1) A single and constant URL 2) A URL with parameters (but no variation in the parameters) for the same dataset	50
Level 4. Web Access unique with parameters to single data Access to information via the web, but allows each data inside the datasets to be accessed individually, or through a shortened URL, or by specific parameters in the query call and includes the date, the version, or the last update. Linked data meet this level and next one Examples 1) url with parameters can address individual data without downloading all the dataset	90
Level 5. API or query language Access to information provides access to specific data of the dataset, either by calling a documented API or through a query language of data sources Examples 1) A SPARQL point to be queried or a documented API (CKAN, socrata, ODS API)	100

2.4.- Data model sharing

Data model in the context of **MELODA** does not analyse the data model itself but the ability of the publisher to share its data model with other entities and provoque a common standard.

In any case if an standard is not RAND or free, punctuation will be level 1.

Description	Weight (%)
Level 1. Not known data model The data model is not released neither explained. Examples 1)	15
Level 2. Own ad hoc data model There are fields designed by the publisher. But the fields are just identified, not explaining the ranges, type and rest of characteristics etc. It is used only by the publisher entity. Examples 1) 2)	35
Level 3. Own ad hoc data model published Although the model is designed by the publisher the specification of the fields of the information is available as a separate information, and it is freely allowed to be used by others. Homemade ontology belong to this level. Examples 1) There is a publish ontology or description freely available	50
Level 4. Local open data model There is a model standardized by a standardization body / local entity (i.e. national) but is poorly adopted (i.e. new standards, standards not widely adopted). Use of a national public ontology belongs to this level. Examples 1)	90
Level 5. Global open data model There is a standardized data model released by a global entity (i.e. supranational, ISO) and/or there is a wide adoption (i.e. when is compulsory by law, etc) Examples 1)	100

2.5.- Geolocated information

Information released could contain information about the location of the information. It does not mean that information has to be geographic but has some fields identifying the location of the information.

Description	Weight (%)
Level 1. No geographic information The information released has not any field referring to the location Examples 1)	15
Level 2. Simple text field The geographic information is just a text field or an own Id that makes difficult to connect this information with other databases. Examples 1) Field name: City, value: Astorga	30
Level 3. Complex text field The geographic information are several fields with text description (the number of a street would be consider as a text). And the fields are hierarchical Examples 1) Field1: Country, value : Spain, Field2: City, value: Astorga	50
Level 4. Coordinates The geographic information include two fields with the coordinates (in some standardized system of coordinates) Examples 1) Field1, Latitude, value :42.605556, Field2 longitude, value: -5.5	90
Level 5. Full geographical information It includes levels 3 and 4 Examples 1) Field1, Latitude, value :42.605556, Field2 longitude, value: -5.5, field 3: Country, value: Spain, field 4: city, value : León	100

In case of several text fields but not clearly creating a unique address it should be considered level 2. When some fields are present but not always filled with values it should be considered as filled unless (never were filled with information)

2.6.- Real-time information

Real time could vary according to the nature of the dataset. While in transport information less than 5 seconds delay could be accepted as real time, in other issues longer periods could be valid.

Description	Weight (%)
Level 1. Longer than week Updating period is longer than a week. 1) The difference between one version and next one is longer than 7 days 2) i.e. monthly, yearly	15
Level 2. Days Updating period ranges from 1 day to 7 days. Examples 1) The difference between one version and next one is longer than 1 day but lower than 7 days 2) Weekly release	40
Level 3. Hours Updating period ranges from 1 hour to 24 hours. Examples 1) The difference between one version and next one is longer than 1 day but lower than 7 days 2) Daily release	70
Level 4. Minutes Updating period ranges from 1 Hour to 1 minute. Examples 1) The difference between one version and next one is longer than 1 minute but lower than 1 hour 2) Hourly release	90
Level 5. Seconds Updating period is lower than 1 minute. It is a dataJet. Examples 1) The difference between one version and next one is lower than 1 minute 2) Updating period measured in seconds	100

In case updating period varies a mean should be calculated.

3.-ASSESSING HOW TO

Data source refers to the dataset to be assessed and **MELODA** 4.10 to the current document. (assessment for 4.10 remains equal to 4.01)

MELODA 4.01 Assessment process
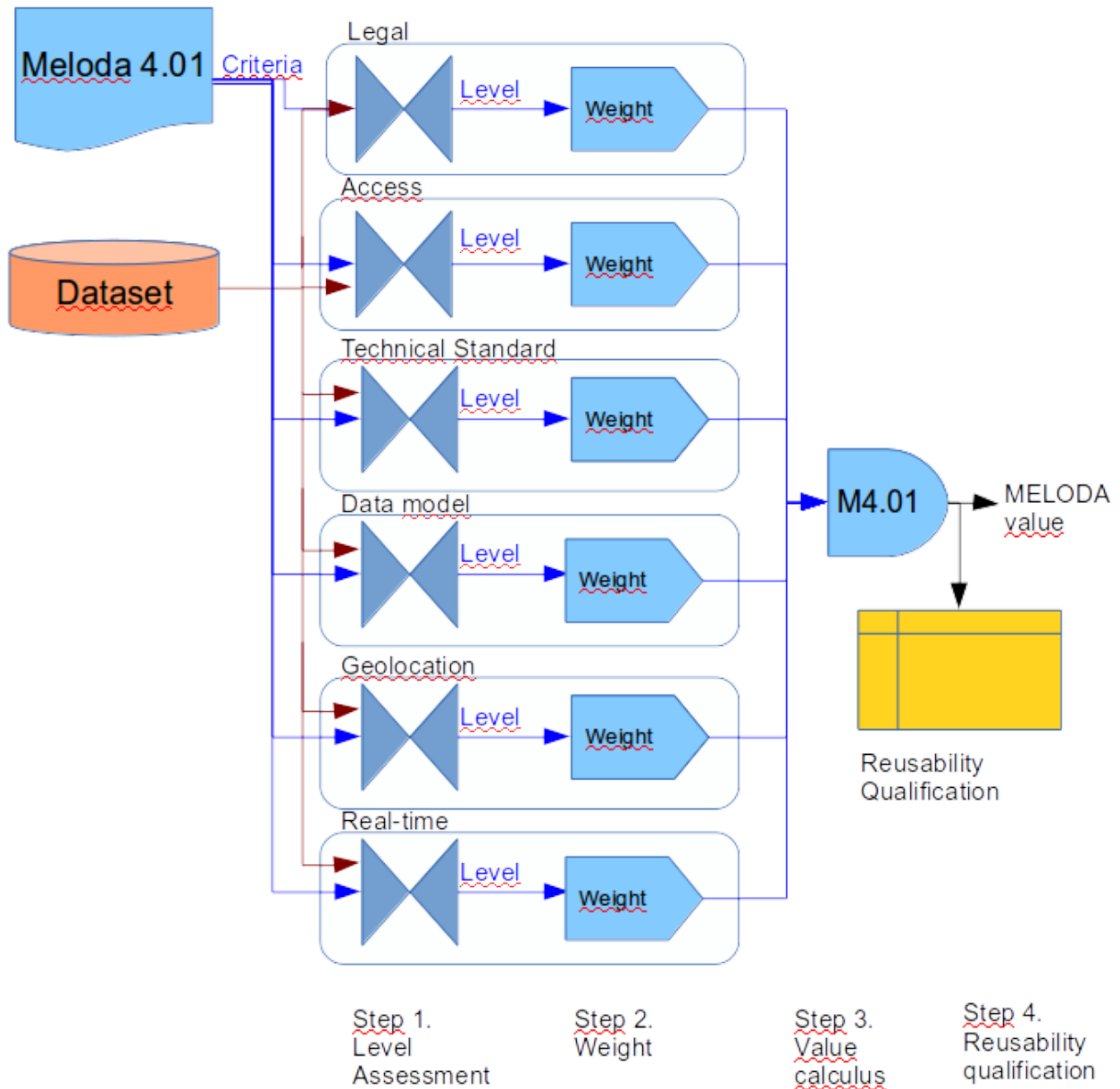
provides an scale ranging from 0 to 100.

3.1.- Example 1.

Institution A releases a dataset which is a spreadsheet in two formats, xls and csv file with a taxonomy of data recently released by the regional government and content can be access to the full dataset through a unique URL. The dataset only contains a field with the address in plain text. It is released monthly but only for private use.

Legal assessing : Level 2 → 10% (private use)

Technical assessing: Level 3 → 60% (csv is an open standard)

Accessibility assessing: Level 3 → 50% (Unique URL to full dataset)

Data model: Level 2 → 35% (Ad hoc model)

Geolocation: Level 2 → 30% (Text Field)

Real Time: Level 1 → 15% (Monthly)

$$100 \cdot \sqrt[6]{10\% \cdot 60\% \cdot 50\% \cdot 35\% \cdot 30\% \cdot 15\%} = 27.9$$

It means *Basic reuse* possible according to global qualification described in section 3.3

You can test with other datasets in this link
http://www.meloda.org/become-assessor/

3.2.- Example 2.

Institution B releases a dataset as linked data (so data has meta data associated) based on a national semantic standard ontology, and allows to be used licensed for commercial re-use and being able to be queried as a SPARQL resource, it includes two fields for the coordinates and it is released as long as data are generated (in seconds).

Legal assessing : Level 4   → 90% (commercial use)

Technical assessing: Level 4   → 100% (Meta data with data)

Accessibility assessing: Level 5   → 100% (Queryable through SPARQL)

Data model: Level 4   → 90% (Local Open Data Model)

Geolocation: Level 4   → 90% (Coordinates without text field)

Real Time: Level 5   → 100% (seconds)

$$100 \cdot \sqrt[6]{90\% \cdot 100\% \cdot 100\% \cdot 90\% \cdot 90\% \cdot 100\%} = 94.9$$

It means *Optimal reuse* possible according to global qualification described in section 3.3

You can test with other datasets in this link
http://www.meloda.org/become-assessor/

http://meloda.org 4.10   12/17

3.3.- Global qualification based on assessment results

Range	Categories	Simple term	Icon
0 to 25	Inadequate for reuse	Limited reuse	
25 to 50	Basic reuse possible	Basic reuse	
50 to 75	Reuse possible but with some improvable characteristic	Good reuse	
75 to 100	The best for reuse	Optimal reuse	

4.- FAQ

Q: The same dataset is released in several standards and licenses. Which should be the final mark?

A: The highest level for every area (legal, technical standard, accessibility, data model, geolocation, timeliness), bearing in mind that this situation could accessed by the user simultaneously.

Q: Could I use MELODA for assessing a full entity publishing several open datasets?

A: Not really but you if you have statistically enough samples from the same entity you could reach a consistent mark.

Q: Why do you create MELODA?

A: Because it is needed a fair indication of the mechanisms for making reusable information (specially for those entities which want to be transparent and they want to impact on their surrounding societies) if people cannot reuse the information the it is useless.
The cost to access information is a way of not being transparent. The lack of a metric discourage the competition and improvement.

Q: Do I need any previous knowledge to use MELODA?

A: Nothing at all. Just to understand how the process of assessing is carried out. You can try it online in this link

Q: I would like to contact with MELODA creators.
A: Just drop an email at alberto.abella@meloda.org

Annex: Theme topics lists based on DCAT-AP

According to standardization of the EU publications office. Reached here.
http://publications.europa.eu/mdr/resource/authority/data-theme/html/data-theme-eng.html

Authority code	Label	Description
AGRI	Agriculture, fisheries, forestry and food	This concept identifies datasets covering such domains as agriculture, fisheries, forestry or food.
ECON	Economy and finance	This concept identifies datasets covering such domains as economy or finance.
EDUC	Education, culture and sport	This concept identifies datasets covering such domains as education, culture or sport.
ENER	Energy	This concept identifies datasets covering the domain of energy.
ENVI	Environment	This concept identifies datasets covering the domain of environment.
GOVE	Government and public sector	This concept identifies datasets covering such domains as government or public sector.
HEAL	Health	This concept identifies datasets covering the domain of health.
INTR	International issues	This concept identifies datasets covering the domain of international issues.
JUST	Justice, legal system and public safety	This concept identifies datasets covering such domains as justice, legal system or public safety.
REGI	Regions and cities	This concept identifies datasets covering such domains as regions or cities.
SOCI	Population and society	This concept identifies datasets covering such domains as population or society.
TECH	Science and technology	This concept identifies datasets covering such domains as science or technology.
TRAN	Transport	This concept identifies datasets covering the domain of transport.

Acknowledgements

Change table

Change	Date	Description
Inclusion of M4J	23-10-11	Meloda for journalists (M4J) include as Annex I
Updata with economical terms to M4J	28-11-11	Meloda for journalists (M4J) include economical terms
Update some terms	30-04-13	Publishing completely
XML and Json	17-5-13	In technical standards
Price	26-6-13	Explicit requirement of free (gratis) access
Requirements	22-8-13	Review of the requirements in order to be more easy to understand Change stages to levels
Assessment and criteria review	06-09-12	
change from the methodology to metric	04-10-13	It is more suitable for what is including. Details on the metric calculations. Standardise from 0 to 10
Detailing criteria	05-10-13	Further explanations on what fulfills a level
Accuracy of data	05-10-13	Introducing the criteria "If data does not represent what the title explains" -> 0
WMS	7-10-13	WMS maximum standard
New version 3.0	6-12-13	Remove meloda for journalists New dimension Data model Process of assessment
data model standards	25-12-13	Details in the new dimension data model
Minor changes	30-4-13	Thanks to and change to commercial uses
Table with usability ranges	09-07-14	A table for assigning a reusability based on the ranges of assessments obtained by a dataset is included
Introducing new dimensions	22-9-14	Introducing new dimensions geographical and time-related.
introducing the standardized topic	22-12-15	Introducing the standardised topic according to the metadata registry of EU. http://publications.europa.eu/mdr/resource/authority/data-theme/html/data-theme-eng.html
Introducing / remarking semantic contents	8-3-2016	
Renamed the goal of the metric and link to the practica	3-5-16	Focused paid on data reusability and linking to a form where a user can get the grade against meloda 4.0
Definitions of dataset and datajet	6-8-16	First version of dataset and dataJet
Reordering terms, logos and fixing mistakes	16-10-16	Moving changes table to the end. Including logos. Fixing mistakes in examples.