|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dimension** | **Sub-dimension** | **Description** | **Variables & Formula (Normalized)**  **Or Reason of exclusion** | |
| **Traceability** | Track of creation | Indicates the presence or absence of metadata associated with the process of creation of a dataset. | s: Source  dc: Date of creation | tc = 2s + dc  tcn = tc/3 |
| Track of updates | Indicates the existence or absence of metadata associated with the updates done to a dataset. | lu: List of updates  du: Dates of updates | tu = lu + du  tun = tu/2 |
| *This metric is set to 0.25 since only the date of the last update is provided in the evaluated portals.* | |
| **Currentness** | Percentage of current rows | Indicates the percentage of rows of a dataset that have current values, it means that they don't have any value that refers to a previous or a following period of time. | Impossible to retrieve resources from a previous period because a versioning module is not implemented in the evaluated portals. | |
| Delay in publication | Indicates the ratio between the delay in the publication (number of days passed between the moment in which the information is available and the publication of the dataset) and the period of time referred by the dataset (week, month, year). | No information on the frequency of updates and the dates of previous updates. | |
| **Expiration** | Delay after expiration | Indicates the ratio between the delay in the publication of a dataset after the expiration of its previous version and the period of time referred by the dataset (week, month, year). |
| **Completeness** | Percentage of complete cells | Indicates the percentage of complete cells in a dataset. It means the cells that are not empty and have a meaningful value assigned (i.e. a value coherent with the domain of the column). | nr: Number of rows nc: Number of columns ic: Number of incomplete cells ncl: Number of cells | ncl = nr\*nc  pcc = (1 – ic/ncl) \* 100  pccn = pcc/100 |
| Percentage of complete rows | Indicates the percentage of complete rows in a dataset. It means the rows that don't have any incomplete cell. | nr: Number of rows nir: Number of incomplete rows | pcpr = (1 – nir/nr) \* 100  pcprn = pcpr/100 |
| **Compliance** | Percentage of standardized columns | Indicates the percentage of standardized columns in a dataset. It just considers the columns that represent some kind of information that has standards associated with it (i.e. geographic information). | Impossible to automatically detect the standardized columns | |
| eGMS Compliance | Indicates the degree to which a dataset follows the e-GMS standard (as far as the basic elements are concerned, it essentially boils down to a specification of which Dublin Core metadata should be supplied) | s: Source dc: Date of creation c: Category t: Title d: Description (*if applicable*) id: Identifier (*if applicable*) pb: Publisher (*if applicable*) cv: Coverage (*recommended only*) l: Language (*recommended only*) | egmsc = s + dc + c + t + 0.25 \* (d + id + pb + l)  egmscn = egmsc/5  *This metric is adjusted because we exclude coverage (since it is only recommended and not available on the portals evaluated).* |
| Five star Open Data | Indicates the level of the 5 star Open Data model in which the dataset is and the advantage offered by this reason. | In this study, the value of this sub-dimension is set to 0.6 (3/5) since we only focus on CSV datasets. | |
| **Understandability** | Percentage of columns with metadata | Indicates the percentage of columns in a dataset that has associated descriptive metadata. This metadata is important because it allows to easily understand the information of the data and the way it is represented. | Unavailability and difficulty in automating the retrieval of metadata columns on the evaluated portals. Since, in some portals, the metadata and the dataset are in the same file and are not well structured. | |
| Percentage of columns in comprehensible format | Indicates the percentage of columns in a dataset that is represented in a format that can be easily understood by the users and it is also machine-readable. | It is impossible to automate this sub-dimension. It requires human intervention to see if a column is understandable or not. | |
| **Accuracy** | Percentage of accurate cells | Indicates the percentage cells in a dataset that has correct values according to the domain and the type of information of the dataset. | nce: Number of cells with errors ncl: Number of cells | pac = (1 – nce/ncl) \* 100  pacn = pac/100 |
| Accuracy in aggregation | Indicates the ratio between the error in aggregation and the scale of data representation. This metric only applies for the datasets that have aggregation columns or when there are two or more datasets referring to the same information but in a different granularity level. | Impossible to automate this sub-dimension. It requires human intervention to check if a column is an aggregation of other columns, since the names of the columns are often ambiguous. | |

*Table 2: List of dimensions and sub-dimensions for the open data quality framework (adapted from (Vetrò et al., 2016)). Gray rows are (sub-) dimensions not taking into account in this study.*