**Mapping of Machine Learning Data and Backend Data**

**Index: sentences**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column Name** | | **Description** | **Values** | **Example values** |
| **Machine Learning** | **Backend** |
| Countries | countries | List of countries of the PCR where the paragraph is extracted | List of countries |  |
| Download Link | urlToFile | Link to download the PCR where the paragraph is extracted | URLs |  |
| File Names | - | File name of the PCR where the paragraph is extracted | string |  |
| Lesson Type | LessonType | Lesson type of the paragraph (if it is a lesson) | "Self Evaluation" |  |
| Month | month | month of the approval date of the PCR where the paragraph is extracted | month |  |
| Project Number | - | Project number of the PCR where the paragraph is extracted | Project numbers |  |
| Sectors | sectors | List of sectors of the PCR where the paragraph is extracted | List of sectors |  |
| Themes | themes | List of themes of the PCR where the paragraph is extracted | List of themes |  |
| Title | title | Title of the PCR where the paragraph is extracted | string |  |
| Year | year | Year of the approval date of the PCR where the paragraph is extracted | year |  |
| annotatedBy | annotatedBy | Author of the annotation made to the paragraph | string |  |
| annotationStatus | annotationStatus | Annotation status of the paragraph | draft, annotated, forReview |  |
| annotationSummary | annotationSummary | Annotation summary made to the paragraph | string |  |
| annotationTitle | annotationTitle | Annotation title made to the paragraph | String |  |
| dataSource | - | Data gathering source of the paragraph (not used) | pcr |  |
| id | id | Primary key of the paragraph | integer |  |
| isLesson | isLesson | Identifies the paragraph if it is a lesson or not | boolean |  |
| lastAnnotated | lastAnnotated | Timestamp of the last annotation done to the paragraph | datetime |  |
| machine generated | summary | Summary provided by the machine learning model to the paragraph | string, blank if not a lesson |  |
| paragraph | paragraph | Extracted paragraph from the PCRs | string |  |
| source | source | Data gathering source of the paragraph | pcr |  |
| reference\_id | reference\_id | Reference id of the paragraph | Padded integer (length: 20) |  |
| related\_lessons | relatedLessons | List of ids of the lessons related to the paragraph (if it is a lesson) | List of ids, empty list if not a lesson |  |
| Modalities | modalities | List of modalities of the PCR where the paragraph is extracted | List of modalities |  |
| Lesson Strength | lessonStrength | Strength of the paragraph (if it is a lesson) | float (0 to 1), -1 if not a lesson |  |
| topic | topic | Topic assigned to the paragraph (if it is a lesson) | integer, -1 if not a lesson |  |
| topTopics | topTopics | List of topics assigned to the paragraph (if it is a lesson) | list of topics, -1 if not a lesson |  |

**Index: topics**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column Name** | | **Description** | **Values** | **Example values** |
| **Machine Learning** | **Backend** |  |  |  |
| Topic Number | key | Primary key | Primary key |  |
| Keywords | keywords | List of keywords that represent the topic | List of keywords |  |
| Keywords Frequencies | frequencies | List of frequencies of the keywords across all paragraphs | List of integers >= 0 |  |
| Number of Lessons | numberOfLessons | Number of lessons assigned to the topic | Integer >= 0 |  |
| Number of PCRs | numberOfPCRs | Number of unique PCRs of all the paragraphs assigned to the topic | Integer >= 0 |  |
| Top Word | topWord | Topic keyword with the highest frequency | string |  |
| x | x | X coordinate assigned to the topic | float (-1 to 1) |  |
| y | y | Y coordinate assigned to the topic | float (-1 to 1) |  |

**Index: pcrs**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column Name** | | **Description** | **Values** | **Example Values** |
| **Machine Learning** | **Backend** |  |  |  |
| id | id | Primary key | Primary key |  |
| Title | title | Title of the PCR | string |  |
| Project Number | projectNumber | Project number of the PCR | string |  |
| Countries | countries | List of countries of the PCR | List of countries |  |
| Sectors | sectors | List of sectors of the PCR | List of sectors |  |
| Themes | themes | List of themes of the PCR | List of themes |  |
| Month Year | monthYear | Approval date of the PCR | string |  |
| Month | month | Month of the approval date of the PCR | string |  |
| Year | year | Year of the approval date of the PCR | string |  |
| Modalities | - | List of modalities of the PCR (website scraped) | List of modalities |  |
| Source Of Funding | sourceOfFunding |  |  |  |
| Safeguard Categories | safeguardCategories |  |  |  |
| Milestone Approval Date | milestoneApprovalDate | Milestone approval date of the PCR | string |  |
| Milestone Signing Date | milestoneSigningDate | Milestone approval date of the PCR | string |  |
| Milestone Effectivity Date | milestoneEffectivityDate | Milestone approval date of the PCR | string |  |
| Download Link | urlToFile | Download link to PCR | URL |  |
| File Names | - | File name of the PCR | string |  |
| MODALITY\_FR\_DUMP | - | List of modalities of the PCR (from dump) | List of modalities |  |
| UNQ\_MODALITY\_FR\_DUMP | modalities | List of unique modalities of the PCR (from dump) | List of modalities |  |