# Appendix A. TAPS Matrix - Questions

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Question** | | | | | | | | | | **ID** | **Component** | **Scope** |
| Who is involved in this project? e.g. researchers, organizations, end users | | | | | | | | | | TA1 | Transparency | Actors |
| Have all the humans involved in the project been identified? This includes  data scientists, downstream users of data science objects, & data subjects (e.g., patients). | | | | | | | | | | TA2 | Transparency | Actors |
| Are there formal agreements or explicitly described relationships between them? Examples of formal agreements: employment contracts (specifically length of contract, percentage of work, employee classifications, & other characteristics that might interact with ACCOUNTABILITY), consent forms, ethical approvals, intellectual property protocols, patents, non-disclosure  agreements, contracts governing engagements | | | | | | | | | | TA3 | Transparency | Actors |
| Are these formal agreements accessible by all of the actors? | | | | | | | | | | TA4 | Transparency | Actors |
| Have the contributions of the humans who are involved been described? | | | | | | | | | | TA5 | Transparency | Actors |
| Are there mechanisms in place to mitigate potential conflicts between actors? | | | | | | | | | | TA6 | Transparency | Actors |
| Is there a provision for continuity and sustainability of the project and its  resources over time? (e.g., do contact emails go to institutional email addresses, not individuals’?) | | | | | | | | | | TA7 | Transparency | Actors |
| Are there any other secondary actors involved or affected by the project? (e.g., investors, funding agency) | | | | | | | | | | TA8 | Transparency | Actors |
| What the outputs or deliverables of the project are? e.g. types of data to  collect/re-use, models or algorithms | | | | | | | | | | TO1 | Transparency | Objects |
| Have all data/metadata/algorithms been described [this may be a link to a  document] | | | | | | | | | | TO1 | Transparency | Objects |
| How will you make data and algorithms you used in the analysis transparent? | | | | | | | | | | TO2 | Transparency | Objects |
| How will you make data and algorithms that you used in the analysis  accessible to others? (such as following the FAIR principles, meeting Open Science guidelines) | | | | | | | | | | TO3 | Transparency | Objects |
| Is any metadata given for the dataset? | | | | | | | | | | TO4 | Transparency | Objects |
| Are your data and metadata in a repository? (provide links such as Github,  Zenodo etc.) | | | | | | | | | | TO5 | Transparency | Objects |
| Are the data and metadata publicly available? | | | | | | | | | | TO6 | Transparency | Objects |
| What is the licence associated with the data? | | | | | | | | | | TO7 | Transparency | Objects |
| Have you measured the quality of the data? If yes, what quality dimensions  and metrics did you use (i.e. accuracy, completeness, consistency, timeliness, uniqueness, and validity )? What is the overall quality of the data? | | | | | | | | | | TO8 | Transparency | Objects |
| What is the quantity of missing values in the data? | | | | | | | | | | TO9 | Transparency | Objects |
| Are there any  data/algorithm? | terms | and | conditions | specified | on | the | use | of | the | TO10 | Transparency | Objects |
| Describe the overall process used to achieve the project’s outputs or deliverables (including data science lifecycle and project planning)? e.g. data  collection, cleaning, data management | | | | | | | | | | TP1 | Transparency | Process |
| How will you collect the data? | | | | | | | | | | TP2 | Transparency | Process |
| How do you clean/preprocess the data? | | | | | | | | | | TP3 | Transparency | Process |
| How do you analyse the data? | | | | | | | | | | TP4 | Transparency | Process |
| How do you maintain/serve the project or your data? | | | | | | | | | | TP5 | Transparency | Process |
| What are the aims of the project? | | | | | | | | | | TI1 | Transparency | Impact |
| What are the outcomes of the project? | | | | | | | | | | TI2 | Transparency | Impact |

|  |  |  |  |
| --- | --- | --- | --- |
| What are the expected impacts of the project? | TI3 | Transparency | Impact |
| Can you describe the implications of the project so that it is understandable  by everyone? | TI4 | Transparency | Impact |
| Can you articulate how this project will affect people's life or research? | TI5 | Transparency | Impact |
| Have you described how this project might be misused? | TI6 | Transparency | Impact |
| Are there any terms and conditions specified on the use of data/algorithm? | TI7 | Transparency | Impact |
| Has the researcher signed a code of conduct agreement? | AA1 | Accountability | Actors |
| Is it clear who is responsible for which part of the outcome? | AA2 | Accountability | Actors |
| Does your organisation have protocols and procedures for dealing with an intentional or unintentional breakdown of transparency,  privacy/confidentiality, or norm breaking? | AA3 | Accountability | Actors |
| Are there protocols and procedures for dealing with conflicts between legal  and ethical positions in projects? | AA4 | Accountability | Actors |
| Does your organisation have a data protection officer? Data ethics review  committees? Data stewards? | AA5 | Accountability | Actors |
| Do employees have access to resources (e.g., training) for doing responsible  data science? | AA6 | Accountability | Actors |
| Do protections exist against lawsuit or damages for individual data scientists,  users, or organisations? | AA7 | Accountability | Actors |
| What is the provenance of the data? | AO1 | Accountability | Objects |
| Do you know the original permissions attached to the data (in the case of  re-use)? | AO2 | Accountability | Objects |
| Are you following the original permissions that are attached to that data? | AO3 | Accountability | Objects |
| Is there a description provided about potential biases, such as confounding  factors? | AO4 | Accountability | Objects |
| Is data quality an issue? If yes, what is the specific issue and how is it  resolved? | AO5 | Accountability | Objects |
| Is there any missing data or missing variables and how are these values  handled? | AO6 | Accountability | Objects |
| Is there a clear, transparent and accessible description of the data collection  and data management processes in the project? | AP1 | Accountability | Process |
| How responsibilities for the processes ( data collection, management or  analysys) are assigned? | AP2 | Accountability | Process |
| Is there any control/checkpoints within the process? | AP3 | Accountability | Process |
| What is the impact or the outcome of the analysis? | AI1 | Accountability | Impact |
| How are the outcomes of the project communicated to the stakeholders? | AI2 | Accountability | Impact |
| How are unintended outcomes communicated to the relevant organisation? | AI3 | Accountability | Impact |
| Are there protocols and procedures for dealing with conflicts between legal  and ethical positions in the project? | AI4 | Accountability | Impact |
| Is there a procedure for dealing with unintended outcomes, either positive or  negative? | AI5 | Accountability | Impact |
| How does your organisation deal with evaluations by external parties about  the responsibility aspects of the project? | AI6 | Accountability | Impact |
| Is there an informed consent form signed by the human subjects in the  project? | PA1 | Privacy | Actors |
| Has a confidentiality agreement or non disclosure agreement been signed (if  applicable)? | PA2 | Privacy | Actors |
| Has a data usage agreement been signed by the researchers involved? | PA3 | Privacy | Actors |

|  |  |  |  |
| --- | --- | --- | --- |
| Has a privacy protection method been used to make the data anonymous or  pseudonymous? | PO1 | Privacy | Objects |
| Are you collecting identifiable personal information? | PO2 | Privacy | Objects |
| Are you collecting sensitive data (as defined by GDPR)? | PO3 | Privacy | Objects |
| Are the data you collect or re-use necessary to answer your research  questions? | PO4 | Privacy | Objects |
| Where/how do you process the data? (do you upload data to some servers to  execute processing? Which tools do you use?) | PO5 | Privacy | Objects |
| Where do you store data during and after your project? | PO6 | Privacy | Objects |
| Is it secure? | PO7 | Privacy | Objects |
| Did you do your data management plan or use a template (such as  https://dmponline.dcc.ac.uk/)? | PO8 | Privacy | Objects |
| Can third parties use the data? if so, how? | PO9 | Privacy | Objects |
| Do your models and results reveal personal information directly? | PO10 | Privacy | Objects |
| Can the outcome be used to identify individuals indirectly? | PO11 | Privacy | Objects |
| Is it possible that other people can combine your model/algorithm with other  resources to identify individuals? | PO12 | Privacy | Objects |
| Did you collect the data from the individuals in question directly, or obtain it  via third parties or other sources (e.g., websites)? | PP1 | Privacy | Process |
| Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise  reproduce, the exact language of the notification itself | PP2 | Privacy | Process |
| Did the individuals in question consent to the collection and use of their data? If so, please describe how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to  which the individuals consented. | PP3 | Privacy | Process |
| If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). | PP4 | Privacy | Process |
| What procedures were used to ensure the anonymity of personal data? For example, removing sensitive data fields, aggregating at a group level, data  encryption mechanisms…ect | PP5 | Privacy | Process |
| Has an information management system been set in place for auditing  purposes? | PP6 | Privacy | Process |
| Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link  or other access point to any supporting documentation | PI1 | Privacy | Impact |
| Does your project have the potential to impact the privacy and confidentiality  of individuals? If yes, how? | PI2 | Privacy | Impact |
| Can your project change society's confidence and trust in data science  projects? If yes, how? | PI3 | Privacy | Impact |
| Will your outcome deliver a positive message to people about how their  privacy and confidentiality have been handled? | PI4 | Privacy | Impact |
| What are the operative social values and norms of the actors that are involved  in the project? (i.e. cultural or ethical norms) | SA1 | Societal Values | Actors |

|  |  |  |  |
| --- | --- | --- | --- |
| In which country(is) or location(s) is the work being conducted (Europe, US,  Asia, etc)? | SA2 | Societal Values | Actors |
| Are there any code of conducts that your project should follow? | SA3 | Societal Values | Actors |
| What are the operative social values and norms that touch on data,  algorithms, and other outputs or deliverables of the project? | SO1 | Societal Values | Objects |
| Are there any laws or regulations in place to secure data protection (like the GDPR, CCPA in California)? | SO2 | Societal Values | Objects |
| What are the practices for sharing research data with the community that your project strives to follow? | SO3 | Societal Values | Objects |
| Is there a data sharing agreement between the partners within or beyond the project? | SO4 | Societal Values | Objects |
| Are there any governing rules affecting your project in regards to use of  devices and/or digital applications? | SO5 | Societal Values | Objects |
| Are there any rules or standards for data collection? | SP1 | Societal Values | Process |
| Did you pre-register your study somewhere (share your research data  management plan)? For example via Center for Open Science (OSF)1. | SP3 | Societal Values | Process |
| Does the data analysis process affect specific population groups? | SP4 | Societal Values | Process |
| If you are collecting data, what are the steps taken to reduce possible societal  bias (i.e gender, population, cultural, age biases) | SP5 | Societal Values | Process |
| If you are using already collected data, what are the steps taken to reduce possible societal biases existing in these data (i.e gender, population, cultural,  age biases)? | SP6 | Societal Values | Process |
| What are the societal impacts of this project? | SI1 | Societal Values | Impact |
| Will the project inform the stakeholders about any possible side-effects or  negative impacts? | SI2 | Societal Values | Impact |
| What could be possible commercial interests in the project’s outcomes? | SI3 | Societal Values | Impact |
| What is the public interest in the project's outcomes once they mature? | SI4 | Societal Values | Impact |
| Is there an identifiable conflict between societal values of the commercial and  public interests in the outcomes? | SI5 | Societal Values | Impact |
| Are there any procedures or norms in place to resolve different interests  between public interest and commercial interests? | SI6 | Societal Values | Impact |
| Did you create a Data protection impact assessment (DPIA)2 which is  obligatory under the GDPR in the Netherlands. | SI7 | Societal Values | Impact |
| Are all of the operative social values and norms in agreement with each  other? If not please explain. | SI8 | Societal Values | Impact |
| Can you describe or estimate the environmental impact of the project (for  example: energy consumption)? | SI9 | Societal Values | Impact |
| Does the project rely on the use of any toxic or non recyclable materials? | SI10 | Societal Values | Impact |

1 Link:

https://osf.io/x5w7h/wiki/06%20LeaderBoard/?hsCtaTracking=2f25be73-35de-4b95-93e8-4f08acc423 32%7Cf2024186-54bb-4626-baab-1785d2f9d71a

2 Link: https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/stcrt-2019-64418.pdf

# Appendix B. The FAIR Guiding Sub-principles

Findability:

* F1. (meta)data are assigned a globally unique and persistent identifier
* F2. data are described with rich metadata (defined by R1 below)
* F3. metadata clearly and explicitly include the identifier of the data it describes
* F4. (meta)data are registered or indexed in a searchable resource Accessibility:
* A1. (meta)data are retrievable by their identifier using a standardized commu- nications protocol
* A1.1 the protocol is open, free, and universally implementable
* A1.2 the protocol allows for an authentication and authorization procedure, where necessary
* A2. metadata are accessible, even when the data are no longer available Interoperability:
* I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
* I2. (meta)data use vocabularies that follow FAIR principles
* I3. (meta)data include qualified references to other (meta)data Reusability:
* R1. meta(data) are richly described with a plurality of accurate and relevant attributes
* R1.1. (meta)data are released with a clear and accessible data usage license
* R1.2. (meta)data are associated with detailed provenance
* R1.3. (meta)data meet domain-relevant community standard

# Appendix C. Datasheets for datasets - Questions

The following is an extract of the questions presented by Gebru et al in the data sheets for datasets paper [15]. We show how we labeled each question in order to map them into the matrix and the questions as stated in Gebru et al work. Some example statements have been shortened. The questions are numbered in the order of appearance within the categories. For example, the first question of the motivation category is labelled 1.1 while the last question in the maintenance category is labelled

7.8 (this category has 8 questions). If questions were accompanied by subsections, they were labelled with letters (i.e 4.1, 4.1.A and 4.1.B). If the subsections were placed on the same matrix cell, then only the main label of the question is used (i.e is 4.1 and all the sub questions belong to the cell transparency/process, then only 4,1 is noted in the mapping, however the sub-questions of 3.7 could not be placed in the same cell as they relate to different scopes, in the matrix, then we noted the exact placement of each sub-question within the matrix). Every section ended with: “Any other comments?”, these questions were labeled but not deemed relevant to be mapped within the matrix.

# Motivation

* + - **1.1** For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a

description.

* + - **1.2** Who created the dataset (e.g., which team) and on behalf of which entity (e.g., institution)?
    - **1.3** Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
    - 1.4 Any other comments? (Not mapped due to being a broad question)

# Composition

* + - **2.1** What do the instances that comprise the dataset represent (e.g., doc- uments, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings)? Please provide a description.
    - **2.2** How many instances are there in total (of each type, if appropriate)?
    - **2.3** Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sam- ple, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representa- tiveness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld).
    - **2.4** What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.
    - **2.5** Is there a label or target associated with each instance? If so, please provide a description.
    - **2.6** Is any information missing from individual instances? If so, please pro- vide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed informa- tion, but might include, e.g., redacted text.
    - **2.7** Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.
    - **2.8** Are there recommended data splits (e.g., training, develop- ment/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
    - **2.9** Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
    - **2.10** Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
    - **2.11** Does the dataset contain data that might be considered confiden-

tial (e.g., data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals’ non-public communications)? If so, please provide a description.

* + - **2.12** Does the dataset contain data that, if viewed directly, might be offen- sive, insulting, threatening, or might otherwise cause anxiety? If so, please

describe why.

* + - 2.13 Does the dataset relate to people? If not, you may skip the remaining questions in this section.
    - **2.13A** Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
    - **2.13B** Is it possible to identify individuals (i.e., one or more natural per- sons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
    - **2.13C** Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; fi- nancial or health data; biometric or genetic data; forms of government iden- tification, such as social security numbers; criminal history)? If so, please provide a description.
    - 2.14 Any other comments? (Not mapped due to being a broad question)

# Collection Process

* + - **3.1** How was the data associated with each instance acquired? Was the data directly observable (e.g., movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of- speech tags)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
    - **3.2** What mechanisms or procedures were used to collect the data (e.g., sensor, manual human curation, software program)? How were these mech- anisms or procedures validated?
    - **3.3** If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabili- ties)?
    - **3.4** Who was involved in the data collection process (e.g., students, crowd- workers) and how were they compensated (e.g., how much were crowd- workers paid)?
    - **3.5** Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time- frame in which the data associated with the instances was created.
    - **3.6** Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any sup- porting documentation.
    - 3.7 Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
    - **3.7A** Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
    - **3.7B** Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
    - **3.7C** Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link

or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

* + - **3.7D** If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
    - **3.7E** Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
    - 3.8 Any other comments? (Not mapped due to being a broad question)

# Preprocessing/Cleaning/Labeling

* + - 4.1 Was any preprocessing/cleaning/labeling of the data done (e.g., dis- cretization or bucketing, tokenization, removal of instances, processing of missing values)? If so, please provide a description.
    - **4.1A** Was the “raw” data saved in addition to the prepro- cessed/cleaned/labeled data (e.g., to support unanticipated uses)? If so, please provide a link or other access point to the “raw” data.
    - **4.1B** Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
    - 4.2 Any other comments? (Not mapped due to being a broad question)

# Uses

* + - **5.1** Has the dataset been used for any tasks already? If so, please provide a description.
    - **5.2** Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
    - **5.3** What (other) tasks could the dataset be used for?
    - **5.4** Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
    - **5.5** Are there tasks for which the dataset should not be used? If so, please provide a description.
    - 5.6 Any other comments? (Not mapped due to being a broad question)

# Distribution

* + - **6.1** Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
    - **6.2** How will the dataset will be distributed (e.g.,website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
    - **6.3** When will the dataset be distributed?
    - **6.4** Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
    - **6.5** Have any third parties imposed IP-based or other restrictions on the

data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any rel- evant licensing terms, as well as any fees associated with these restrictions.

* + - **6.6** Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
    - 6.7 Any other comments? (Not mapped due to being a broad question)

# Maintenance

* + - **7.1** Who is supporting/hosting/maintaining the dataset?
    - **7.2** How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
    - **7.3** Is there an erratum? If so, please provide a link or other access point.
    - **7.4** Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
    - **7.5** If the dataset relates to people, are there applicable limits on the re- tention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
    - **7.6** Will older versions of the dataset continue to be sup- ported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
    - **7.7** If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
    - 7.8 Any other comments? (Not mapped due to being a broad question)