Plan Overview

A Data Management Plan created using DMPonline

Title: My plan (Horizon 2020 DMP)

Creator: Antonio Sgorbissa

Principal Investigator: Antonio Sgorbissa

Data Manager: Antonio Sgorbissa

Affiliation: Other

Funder: European Commission

Template: Horizon 2020 DMP

ORCID iD: 0000-0001-7789-4311

Project abstract:

The groundbreaking objective of CARESSES is to build culturally competent care robots, able to autonomously re-configure their way of acting and speaking, when offering a service, to match the culture, customs and etiquette of the person they are assisting. By designing robots that are more sensitive to the user's needs, CARESSES' innovative solution will offer elderly clients a safe, reliable and intuitive system to foster their independence and autonomy, with a greater impact on quality of life, a reduced caregiver burden, and an improved efficiency and efficacy. The need for cultural competence has been deeply investigated in the Nursing literature. However, it has been totally neglected in Robotics. CARESSES stems from the consideration that cultural competence is crucial for care robots as it is for human caregivers. From the user's perspective, a culturally appropriate behavior is key to improve acceptability; from the commercial perspective, it will open new avenues for marketing robots across different countries. CARESSES will adopt the following approach. First, we will study how to represent cultural models, how to use these models in sensing, planning and acting, and how to acquire them. Second, we will consider three (physically identical) replicas of a commercial robot on the market and integrate cultural models into them, by making them culturally competent. Third, we will test the three robots, customized for three different cultures, in the EU (two cultural groups) and Japan (one cultural group), on a number of elderly volunteers and their informal caregivers. Evaluation will be conducted through quantitative and qualitative investigation. To achieve its groundbreaking objective, CARESSES will involve a multidisciplinary team of EU and Japanese researchers with a background in Transcultural Nursing, AI, Robotics, Testing and evaluations of health-care technology, a worldwide leading company in Robotics and a network of Nursing care homes.

ID: 14907

Last modified: 27-03-2021

Grant number / URL: 737858

My plan (Horizon 2020 DMP) - Initial DMP

1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

Three datasets have been selected to be included in the Data Management Plan:

- Dataset 1: Cultural Knowledge Base (CKB)
- Dataset 2: Interaction logs (IL)
- Dataset 3: End-Users Responses (EUR)

Dataset 1: Cultural Knowledge Base (CKB)

State the purpose of the data collection/generation

The purpose of WP1 and WP2 is to: 1) collect the corpus of knowledge allowing an assistive robot to exhibit a culturally competent behavior (with a specific focus on the three cultures considered during the final testing stage); and 2) formalize it in a framework allowing for the automated acquisition, update and retrieval of culture-related information. This framework is the Cultural Knowledge Base, that will allow for performing a cultural assessment of the user and aligning plans and sensorimotor behaviours to the user's cultural identity.

Explain the relation to the objectives of the project

The design and development of a framework for cultural knowledge representation, allowing for the automated acquisition, update and retrieval of culture-related information is the purpose of KRA2, and directly matching the scientific objectives O2, O3, O4 and the technological objectives O5, O6 of the project. Moreover, the Cultural Knowledge Base is key to performing a cultural assessment of the user and aligning plans and sensorimotor behaviours to the user's cultural identity, which is the main goal of the project.

Specify the types and formats of data generated/collected

- *What format will your data be in (SPSS, Open Document Format, tab-delimited format, etc)?*

The CKB will be an ontology written in the OWL 2 language (https://www.w3.org/OWL/).

- *Why have you chosen to use a particular format?*

OWL is described as "a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things. OWL is a computational logic-based language such that knowledge expressed in OWL can be exploited by computer programs, e.g., to verify the consistency of that knowledge or to make implicit knowledge explicit." (https://www.w3.org/OWL/) As such, the language perfectly matches the requirements for the Cultural Knowledge Base, as described in the previous sections.

- *Do the chosen formats and software enable sharing and long-term validity of data?*

OWL and its current version OWL 2 are a standard developed by the W3C consortium (http://www.w3.org/), which is the main international standards organization for the World Wide Web, and are arguably the most popular knowledge representation language. OWL (OWL 2 since 2009) was first published in 2004 and it has always been actively maintained by the W3C.

Specify if existing data is being re-used

- *Are there any existing data or methods that you can reuse?*

We will reuse, as far as our application permits it, existing ontologies for the description of concepts of relevance in the context of the CARESSES project.

- *Do you need to pay to reuse existing data?*

Many ontologies are published under licenses that allow for free use, sharing and reuse, such as the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). At the moment, there is no evidence that we will have to reuse data which is not

freely accessible.

- *Are there any restrictions on the reuse of third-party data?*

We will refer to the licenses of the third-party ontologies we will include in the CKB ontology (if any) to determine possible restrictions.

- *Can the data that you create - which may be derived from third-party data - be shared?*

We will refer to the licenses of the third-party ontologies we will include in the CKB ontology (if any) to define the conditions under which the CKB can be accessed, used and shared.

Specify the origin of the data

- *How are the data produced and collected (possibly with reference to the CARESSES WorkPlan)?*

Task 1.1, Task 1.2 and Task 1.3 are devoted to the identification, collection and validation of all the knowledge required by culturally competent robots for elderly assistance. At the same time, Task 2.1, Task 2.2 and Task 2.3 are devoted to the identification and development of the framework and tools for the representation of cultural knowledge. Task 1.4 is devoted to the formalization of the knowledge collected in Tasks 1.1-3 with the tools developed in Tasks 2.1-3

State the expected size of the data

- *State the expected size, not necessarily in terms of "memory storage"; this can be the number of records in a Database, a number of "facts" or "rules", values versus time, and so on.*

Ontologies are usually described in terms of number of classes, properties, datatypes and instances they provide (see for example the Time Ontology: http://lov.okfn.org/dataset/lov/vocabs/time). As a reference, the DOGONT ontology for the description of intelligent domotic environments (http://lov.okfn.org/dataset/lov/vocabs/dogont) describes 893 classes and 74 properties.

Outline the data utility: to whom it will be useful

An ontology describing the corpus of knowledge required for culturally competent assistive robots can be useful: 1) in the field of Robotics, as a guideline and reference for the development of robots able to interact with people while keeping cultural information into account; 2) in the field of Transcultural Nursing, as a validated and publicly available ontology for the description of concepts related to cultural competence and the detailing of a number of cultures (specifically, the ones to be considered during the testing phase of CARESSES).

Please provide a concrete example of the data produced in the right format

Example 1: OWL ontology (with examples of object properties, data properties, classes and individuals) describing some of the concepts contained in the CKB

```
<?xml version="1.0"?>
<rdf:RDF xmlns="http://example.com/caressesontology#"
     xml:base="http://example.com/caressesontology"
    [...]
     xmlns:caressesontology="http://example.com/caressesontology#">
    <owl:Ontology rdf:about="http://example.com/caressesontology">
        <rdfs:comment>This is the Knowledge Base for Caresses</rdfs:comment>
    </owl:Ontology>
    <!--
    ///////////////////////////////////////////////////////////////////////////////////
    // Object Properties
    ///////////////////////////////////////////////////////////////////////////////////
     -->
    <!-- http://example.com/caressesontology#has_Positive -->
    <owl:ObjectProperty rdf:about="http://example.com/caressesontology#has_Positive">
        <rdfs:domain rdf:resource="http://example.com/caressesontology#User"/>
        <rdfs:range rdf:resource="http://example.com/caressesontology#Topic"/>
    </owl:ObjectProperty>
    <!--
    ///////////////////////////////////////////////////////////////////////////////////
    // Data properties
    ///////////////////////////////////////////////////////////////////////////////////
     -->
    <!-- http://example.com/caressesontology#age -->
    <owl:DatatypeProperty rdf:about="http://example.com/caressesontology#age">
        <rdfs:domain rdf:resource="http://example.com/caressesontology#User"/>
        <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#int"/>
```

```
    </owl:DatatypeProperty>
    <!-- http://example.com/caressesontology#gender -->
    <owl:DatatypeProperty rdf:about="http://example.com/caressesontology#gender">
        <rdfs:domain rdf:resource="http://example.com/caressesontology#User"/>
        <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
<!--
    ///////////////////////////////////////////////////////////////////////////////////
    // Classes
    ///////////////////////////////////////////////////////////////////////////////////
     -->
    <!-- http://example.com/caressesontology#AlmondChicken -->
    <owl:Class rdf:about="http://example.com/caressesontology#AlmondChicken">
        <rdfs:subClassOf rdf:resource="http://example.com/caressesontology#ChineseFood"/>
        <owl:disjointWith rdf:resource="http://example.com/caressesontology#CantoneseFriedRice"/>
        <rdfs:comment>Almond chicken</rdfs:comment>
    </owl:Class>
    <!-- http://example.com/caressesontology#Badminton -->
    <owl:Class rdf:about="http://example.com/caressesontology#Badminton">
        <rdfs:subClassOf rdf:resource="http://example.com/caressesontology#Sport"/>
        <rdfs:comment>Badminton</rdfs:comment>
    </owl:Class>
<!--
    ///////////////////////////////////////////////////////////////////////////////////
    // Individuals
    ///////////////////////////////////////////////////////////////////////////////////
     -->
    <!-- http://example.com/caressesontology#SCH_AlmondChicken -->
    <owl:NamedIndividual rdf:about="http://example.com/caressesontology#SCH_AlmondChicken">
        <rdf:type rdf:resource="http://example.com/caressesontology#AlmondChicken"/>
        <likeliness rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">0.7</likeliness>
        <neg>I really don’t like almond chicken</neg>
        <pos>Almond chicken is my favourite chinese food!</pos>
        <pos>Almond chicken is so tasty!</pos>
        <pos>Chinese almond chicken is lovely!</pos>
        <pos>I always eat almond chicken</pos>
        <pos>I love almond chicken!</pos>
        <poswait>Almond chicken is delicious, isn’t it?</poswait>
        <poswait>Do you know a good place here around where I can eat almond chicken?</poswait>
        <poswait>Have you eaten almond chicken recently?</poswait>
        <poswait>What about some almond chicken today?</poswait>
        <que>Do you like almond chicken?</que>
        <topicname>AlmondChicken</topicname>
        <rdfs:comment>Topic Almond Chicken related to a Chinese User</rdfs:comment>
    </owl:NamedIndividual>
```

Dataset 2: Interaction Logs (IL)

State the purpose of the data collection/generation

The IL data set is the collection of messages shared among the CARESSES components during interactions between the culturally competent robot and a person. Each IL file captures the events occurred during the encounter, the actions and status of the person (as perceived by the robot) and the actions of the robot, and it is acquired to the aim of allowing offline analyses and replays of the events occurred during the interaction.

Explain the relation to the objectives of the project

In the course of Task 5.6, the analysis of the Interaction Logs is key to evaluate the performance of the components developed in WP2, WP3 and WP4, which refer to the technical objectives O5-O12 of the project. In the context of the end-user evaluation performed in WP7, the analysis of the Interaction Logs collected during the tests in WP6 can help in assessing the performance of the culturally competent assistive robot, which contributes to the validation objective O15.

Specify the types and formats of data generated/collected

- *What format will your data be in (SPSS, Open Document Format, tab-delimited format, etc)?*

The IL data set will be a collection of text files in CSV format, which is among the most readable formats for information storage. Each line corresponds to a record, i.e. all the info related to a message shared over universAAL by any of the software components of the culturally competent robot during an encounter with a person. A record is divided into fields, separated by a delimiter (e.g. a comma). Fields of relevance in our case include: 1) timestamp of the message; 2) owner of the message; 3) content of the message.

- *Why have you chosen to use a particular format?*

The CSV format is a very popular format for data exchange, widely supported by consumer, business and scientific applications (e.g Microsoft Excel, MATLAB). The fields to store in the IL records comply with popular standards for log files (e.g. the ROS Bag file format for the log files of ROS applications defined in http://wiki.ros.org/Bags/Format/2.0, or the Extended Log file Format for the log files of web servers defined in http://www.w3.org/TR/WD-logfile.html). In particular, the ROS middleware (http://www.ros.org/) is the de-facto standard in robotics applications and log files written in the ROS Bag file format can be replayed and accessed within ROS by any other component. Conversion from the CSV format to the ROS Bag file format is not difficult (see http://answers.ros.org/question/119211/creating-a-ros-bag-file-from-csv-file-data/).

- *Do the chosen formats and software enable sharing and long-term validity of data?*

The CSV format, is among the most readable formats for information storage, supported by the vast majority of software for numerical and data analysis.

Specify if existing data is being re-used

- *Are there any existing data or methods that you can reuse?*

No. The IL data will be entirely produced in the course of CARESSES, during interactions between the culturally competent robot and a person.

- *Can the data that you create - which may be derived from third-party data - be shared?*

We do not foresee any restriction to sharing the IL data set.

Specify the origin of the data

- *How are the data produced and collected (possibly with reference to the CARESSES WorkPlan*

Interaction Logs are collected during two separate stages of the project: 1) in the course of Task 5.6 (evaluation of the integrated CARESSES modules as validation stage within the development process) and 2) in the course of Task 6.3 (experimental evaluation of the culturally competent robot with the participants in the control and experimental groups) and Task 6.4 (experimental evaluation of the culturally competent robot in the smart house iHouse).

State the expected size of the data

- *State the expected size, not necessarily in terms of "memory storage"; this can be the number of records in a Database, a number of "facts" or "rules", values versus time, and so on.*

The IL data set will be described in terms of number of files (i.e., number of recorded interactions between the culturally competent robot and a person) and number of records in each file.

Outline the data utility: to whom it will be useful

The IL data set, as a collection of quantitative data describing interactions between a person and an assistive robot can be useful to academic and industrial researchers aiming at defining guidelines, best practices and standards in the field of Human-Robot Interaction (e.g., identifying which robot actions are frequently requested by people, identifying recurrent sequences of robot actions – human actions that a robot could rely on to exhibit predictive behaviours). Portions of the dataset may also be used by roboticists for the development and testing of specific robotic applications (e.g., the IL dataset can be used to train and test algorithms for learning the habits/routines of a person from the analysis of recurring events).

Please provide a concrete example of the data produced in the right format
Log of messages shared over universAAL, as provided by the universAAL component Log Monitor

Message Type	Timestamp	Content
D5.1 *(user request)*	1496049951891	[Remind_medication : blue_pill : between 12.00 and 12.30]
D6.1 *(user state)*	1496049973842	[Greta : Greta Ahlgren : 10/04/2017: 12:05 : (2.3, 1.0, 0.0): (1.2, 0.0, 90.0) : Kitchen.FridgeArea : Standing : - : Cooking : Eating : Excited]

Dataset 3: End-Users Responses (EUR)

State the purpose of the data collection/generation

The end-user evaluation of the culturally competent robot performed within the CARESSES project implies gathering the responses

of the end user to a number of tools (at present they include: Adapted CCA tool, SF-36, ZBI, QUIS) and the transcripts of qualitative semi-structured interviews. The analysis of such responses: 1) enables us to be able to describe the differences in baseline characteristics of the clients within and between the arms they are allocated to, which is crucial for controlling and thus minimizing the impact of confounding variables; 2) allows for assessing the impact of the (culturally competent) assistive robot in terms of quality of life, increased independence and autonomy, health and care efficiency gains.

Explain the relation to the objectives of the project

The assessment of the impact of the culturally competent assistive robot on the lives of elderly people and their informal carers is a key goal of the whole CARESSES project. More specifically, the evaluation of the robot with elderly participants belonging to different cultures refers to validation objectives O15, O16 and O17.

Specify the types and formats of data generated/collected

- *What format will your data be in (SPSS, Open Document Format, tab-delimited format, etc)?*

Quantitative data collected from structured questionnaires will comply with the SPSS v21 format. Qualitative data collected from semi-structured interviews will be transcribed verbatim using Microsoft Word and subsequently imported into QSR NVivo 11.

- *Why have you chosen to use a particular format?*

We hold expertise in both SPSS and QSR NVivo, both of which are advanced and appropriate analytical software tools.

- *Do the chosen formats and software enable sharing and long-term validity of data?*

Yes.

Specify if exsisting data is being re-used

- *Are there any existing data or methods that you can reuse?*

No. The EUR data will be entirely produced in the course of CARESSES, during interactions between the culturally competent robot and the end-users recruited for the testing phase.

- *Do you need to pay to reuse existing data?*

No, but we will need permission to use outcome tools of interest.

- *Are there any restrictions on the reuse of third-party data?*

No.

- *Can the data that you create - which may be derived from third-party data - be shared?*

Yes within the CARESSES consortium. Anonymised / non-identifiable data will be used in outputs.

Specify the origin of the data

- *How are the data produced and collected (possibly with reference to the CARESSES WorkPlan*

Quantitative data will be produced in the course of Tasks 6.1, 6.2, 6.3, 7.1, 7.3 through the following structured tools applied during the testing phase:
- Background data: Cultural group, age, gender, client diagnosis, educational level, marital status, religion and religiosity, and data collected during screening (i.e. aggression, cognitive competence etc)
- Outcome data: Adapted RCTSH Cultural Competence Assessment Tool (CCATool, Papadopoulos et al., 2004), Short Form (36) Health Survey (SF-36 v2, Hays et al 1993), the Zarit Burden Inventory (ZBI; Zarit et al., 1980), and Questionnaire for user interface satisfaction (QUIS) (Chin et al, 1988). Also we need to record screening results and response rates – all of this data will be compiled into SPSS.
Qualitative data will be collected in the course of Tasks 6.1, 6.2, 6.3, 7.2, 7.3 during semi-structured interviews with clients and informal caregivers.

State the expected size of the data

- *State the expected size, not necessarily in terms of "memory storage"; this can be the number of records in a Database, a number of "facts" or "rules", values versus time, and so on.*

SPSS database: 45 clients and up to 45 caregivers, background and screening associated data per client, background data per caregiver, two time points for SF 36 and ZBI, one time point for CCATool and QUIS. Therefore, approximately 90 rows and 100 columns of data
NVivo database: Transcripts of 15 clients and up to 15 caregivers

Outline the data utility: to whom it will be useful

Anyone involved with the analysis and dissemination activities associated with WP7 data.

Please provide a concrete example of the data produced in the right format

SPSS data:

Client number, cultural group, age, gender, diagnosis, educational level, marital status, religion, religiosity, InterRai aggression, InterRai cognitive competence, CCATool questions and scores, SF36 questions and scores, ZBI questions and scores, QUIS questions and scores

1	*WE*	*71*	*M*	*Mild dementia*	*University degree*	*Widowed*	*C of E*	*medium*	*low*	*high*	*5*	*...*
2	*IND*	*77*	*F*	*Depression*	*College level*	*Widowed*	*Hinduisim*	*low*	*low*	*high*	*7*	*...*

2. FAIR data

2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

Dataset 1: Cultural Knowledge Base (CKB)

Outline the discoverability of data (metadata provision)

- *What metadata, documentation or other supporting material should accompany the data for it to be interpreted correctly?*

The Linked Open Vocabularies (LOV) initiative (http://lov.okfn.org/dataset/lov) hosts a large number of vocabularies and ontologies for the semantic web, and actively promotes the design and publication of high quality ontologies. Their recommendations for the metadata and documentation supporting an ontology are publicly available at http://lov.okfn.org/Recommendations_Vocabulary_Design.pdf. We will adhere, as far as the peculiarities of our application allow it, to those guidelines in the preparation of the metadata and documentation of the CKB. As an example, the above recommendations define the fields and formats of the metadata to associate to classes and properties as “rdfs:label” (element title), “rdfs:comment” (element role), “rdfs:isDefinedBy” (explicit link between an element and the namespace it belongs to), “vs:term_status” (element status among “stable”, “testing”, “unstable”, “deprecated”).

The ontology itself together with the metadata allow for the automatic generation of documentation.

We will also provide as much as possible of the original cultural information formalized in the CKB, to provide the rationale for the formalization we propose and foster the research on, on the one hand, what knowledge makes for a culturally competent robot and, on the other hand, how such knowledge should be formalized for its effective use by the robot.

- *What information needs to be retained to enable the data to be read and interpreted in the future?*

The metadata written in accordance with the aforementioned recommendations and the documentation automatically generated from the CKB ontology contain all the information to be retained to ensure its readability.

- *How will you capture / create the metadata?*

Metadata will be created and updated manually, concurrently with the data, in the course of WP1 and WP2 as described in the above sections. The creation/update of metadata, specifically consists in the writing of a number of text fields for each element of the ontology.

- *Can any of this information be created automatically?*

Metadata will be manually inserted in the CKB ontology. A number of tools exist to automatically generate the documentation of an ontology starting from its description and metadata in the OWL / RDF language, e.g. Parrot http://idi.fundacionctic.org/parrot/parrot.

- *What metadata standards will you use and why?*

We will adhere to the recommendations for metadata and documentation of ontologies drafted by the LOV initiative (publicly available at http://lov.okfn.org/Recommendations_Vocabulary_Design.pdf), which are aimed at maximizing the readability and usability of the ontology by other users.

Such guidelines require no big effort for the production of the metadata and documentation and ensure the compatibility with the requirements of many freely available tools, such as WebVOWL, for the interpretation and visualization of ontologies (for example in

the case of the Time ontology http://visualdataweb.de/webvowl/#iri=http://www.w3.org/2006/time).

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?
Once made publicly available, the CKB will be identified by a unique Uniform Resource Identifier (URI). We will consider suggesting the CKB ontology for inclusion in databases such as Protégé Ontology Library (https://protegewiki.stanford.edu/wiki/Protege_Ontology_Library) and LOV, which provides rich indexing and search tools.

Outline naming conventions used
A number of different style guidelines and naming conventions for ontologies have been proposed. [1] surveys the most popular ones and tries to extrapolate guidelines which are valid in a multilingual scenario. Considering the intrinsic multilingual nature of the CARESSES project, we will adopt, whenever possible, the guidelines they propose for multilingual applications.
[1] Montiel-Ponsoda, E., Vila Suero, D., Villazón-Terrazas, B., Dunsire, G., Escolano Rodríguez, E., & Gómez-Pérez, A. (2011). Style guidelines for naming and labeling ontologies in the multilingual web.

Outline the approach towards search keyword
Indexing and search engines automatically identify the names of classes, properties, datatypes and instances as valid search keywords (see for example http://lov.okfn.org/dataset/lov/about).

Outline the approach for clear versioning
The metadata of the CKB ontology will include information about the date of publication of the ontology (“dc:issued” element of the Dublic Core vocabulary for resource description), date of the last modification (“dc:modified”), version code (“owl:versionInfo”) and change log with respect to the previous version (rdfs:comment). In addition to this, popular Git repository hosting services provide a large number of tools for version control.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how
We will adhere, as far as the peculiarities of our project will allow it, to the recommendations for metadata and documentation of ontologies drafted by the LOV initiative (http://lov.okfn.org/Recommendations_Vocabulary_Design.pdf).

Dataset 2: Interaction Logs (IL)

Outline the discoverability of data (metadata provision)

- *What metadata, documentation or other supporting material should accompany the data for it to be interpreted correctly?*

The IL data set requires documentation describing: 1) the system’s functional architecture (in terms of what the different CARESSES components require and provide and how they are connected) and 2) the details of the messages shared over universAAL, that are stored in the IL files.
Metadata are divided in two categories. Metadata related to a IL file (e.g. time and location of the recorded interaction) will be manually added to each file. Metadata related to the messages (time, owner, message type) are stored in the fields of each record together with the message content and will be automatically associated to the messages by the logging tool.

- *What information needs to be retained to enable the data to be read and interpreted in the future?*

The documentation and metadata written in accordance with the aforementioned specifications contain all the information to be retained to ensure the readability of the IL files.

- *How will you capture / create the metadata?*

IL files are automatically generated during an encounter between the culturally competent robot and a person. As mentioned above, metadata related to the messages (time, owner, message type) will be automatically created by the universAAL communication middleware and stored in the IL files at runtime by the logging tool. Metadata related to a IL file (e.g. time and location of the recorded interaction) will be manually added at a later stage.

- *Can any of this information be created automatically?*

Metadata related to the messages are automatically created by the universAAL communication middleware. Some of the metadata related to a IL file (e.g. starting time and location of the recorded interaction) can also be generated automatically by the logging tool.
Documentation cannot be generated automatically.

- *What metadata standards will you use and why?*

The rationale for choosing the metadata related to the messages, stored in the fields of each record together with the message content, draws inspiration from popular standards for log files (e.g. the ROS Bag file format for the log files of ROS applications defined in http://wiki.ros.org/Bags/Format/2.0, or the Extended Log file Format for the log files of web servers defined in http://www.w3.org/TR/WD-logfile.html). The notation and naming convention will adhere with those of the universAAL platform.

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique

identifiers such as Digital Object Identifiers?
Github (https://github.com/) is among the largest and most popular repository hosting services. Github repositories can be given a DOI and released using the data archiving tool Zenodo (https://zenodo.org/), which also ensures that all metadata required for the identification of the repository are filled before its public release. We will consider this option for the publication of the IL dataset.

Outline naming conventions used
The metadata associated with the dataset itself with adhere to the conventions of the chosen archiving tool (e.g., Zenodo). Metadata associated with files and records with follow the naming convention of the universAAL platform.

Outline the approach towards search keyword
Archiving services such as Zenodo allow for specifying a list of search keywords to associate with the dataset, as part of the publication process.

Outline the approach for clear versioning
Github (as most repository hosting services) provides a large number of tools for version control, in particular allowing for making different releases of a repository. By default, Zenodo takes an archive of the associated Github repository every time a new release is created.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how
The metadata requested by Zenodo for the publication of an archive comply with several standard metadata format such as MARCXML, Dublin Core and DataCite Metadata Schema (http://about.zenodo.org/policies/).

Dataset 3: End-Users Responses (EUR)

Outline the discoverability of data (metadata provision)

- *What metadata, documentation or other supporting material should accompany the data for it to be interpreted correctly?*

The EUR data set requires metadata describing the real-world meaning of values, variables and files, as well as technical information such as variable types and formats. Qualitative metadata pertaining to the file type, data source, the geographic and temporal coverage, source descriptions, annotations, coding structures and explanations will be documented,

- *What information needs to be retained to enable the data to be read and interpreted in the future?*

The metadata written in accordance with the aforementioned specifications contain all the information to be retained to ensure the readability of the EUR data.

- *How will you capture / create the metadata?*

SPSS stores all metadata associated with a dataset in a Dictionary, and provides tools for its creation, validation and export in easily readable formats. The SPSS Dictionary will be created together with the insertion of the quantitative EUR data in SPSS. QSR NVivo 11 also enables data management including providing tools for documentation files, classification and attributes, and enables exporting into a wide range of formats appropriate for archiving.

- *Can any of this information be created automatically?*

SPSS metadata to be stored in the Dictionary will be created manually. For NVivo, a log of information about the data sources, editing done, coding and analysis carried out is created automatically. Other information will be created manually.

- *What metadata standards will you use and why?*

Metadata and documentation standards will adhere to those described by the UK Data Archive (http://www.data-archive.ac.uk/), which is the UK's largest collection of digital research data in the social sciences and humanities and is connected to a network of data archives across the world.

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?
The UK Data Archive supports the use of persistent identifiers across its work so that data, metadata and other outputs can be reliably referenced and linked, in particular promoting the association of data sets with the ORCID of the contributors and with DataCite DOIs for persistent data citation. Other archives, such as Zenodo (https://zenodo.org/), allow for associating a DOI to the data sets. We will consider these options for the publication of the EUR data set.

Outline naming conventions used
The EUR data set will adhere, as far as possible, to the conventions of the chosen archiving service (e.g., UK Data Archive, Zenodo) and of the tools it refers to.

Outline the approach towards search keyword
Both aforementioned archiving services make sure that search keywords and metadata required for finding the data set with their

search tools are provided as part of the publication process.

Outline the approach for clear versioning
A number of solutions for clear versioning of the EUR data set are available. Most metadata standards (e.g. Dublin Core) allow for specifying the version and other related information inside the data set. Moreover, a number of repository hosting services (e.g. Github) provide a large number of tools for version control, in particular allowing for making different releases of a repository.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how
Metadata creation of the quantitative and qualitative data will adhere to the standards described by the UK Data Archive. It is important to mention that the UK Data Archive is a member of the Data Documentation Initiative, whose aims include the development of robust metadata standards for social science data.

2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

Dataset 1: Cultural Knowledge Base (CKB)

Is the data made openly available (YES/NO/PARTIALLY)
YES

Specify how and where (in which repository) the data will be made available
For the storage of the CKB ontology we will consider different solutions, evaluating their performance in terms of data persistence, security and accessibility. To allow other researchers to easily find the ontology, we will apply for its insertion in popular ontology libraries and search engines, such as Protégé Ontology Library, LOV and Google.

Specify what methods or software tools are needed to access the data?

- *Name the required methods or software tools*

A list of existing tools for accessing, visualizing and managing ontologies such as the CKB ontology is available at: https://en.wikipedia.org/wiki/Ontology_(information_science)#Editor

- *Is the software pre-existing or developed as an output of CARESSES*

All the tools listed above are pre-existing and independent from CARESSES.

- *Is documentation about the software available to access the data included?*

Most of the tools listed above provide documentation and support (see for example Protégé: http://protege.stanford.edu/)

- *Is it possible to include the relevant software (e.g. in open source code)?*

Many of the tools listed above (e.g. Protégé) are open source.

Dataset 2: Interaction Logs (IL)

Is the data made openly available (YES/NO/PARTIALLY)
YES

Specify how and where (in which repository) the data will be made available
We are considering to publish the IL dataset on a public Github repository and to use Zenodo for assigning it a DOI.

Specify what methods or software tools are needed to access the data?

- *Name the required methods or software tools*

Files in the CSV format can be accessed by a wide variety of software applications, including proprietary (e.g., Microsoft Excel, MATLAB) and open source applications (e.g. Open Office Calc, Octave, R).

- *Is the software pre-existing or developed as an output of CARESSES*

All the applications mentioned above are pre-existing and independent from CARESSES.
According to the needs of the project, we will maybe develop software applications (e.g. ROS packages) or scripts for existing software (e.g. MATLAB or R scripts) specifically for the management of the data within the IL files. In such case, we will consider publishing such code together with the dataset.

- *Is documentation about the software available to access the data included?*

Most of the applications mentioned above come with rich documentation and support functionalities (see for example MATLAB https://uk.mathworks.com/support/?s_tid=gn_supp).

- *Is it possible to include the relevant software (e.g. in open source code)?*

The applications mentioned above which are not open source (e.g. Microsoft Excel, MATLAB) provide free trials. Moreover, both Microsoft and Mathworks have special licensing contracts for students and academic institutions.

Dataset 3: End-Users Responses (EUR)

Is the data made openly available (YES/NO/PARTIALLY)
PARTIALLY

- *If some data is kept closed provide rationale for doing so*

We will withhold screening data (pertaining to cognitive competence and aggression) since this data is purely used to determine their eligibility rather than for data analysis.

- *With whom will you share the data, and under what conditions?*

This data will not be shared unless we consider the data to be of considerable health importance to the research participant. In this case we shall be guided by our incidental findings policy and may in some cases disclose this data to the research participant.

- *Specify how and where (in which repository) the data will be made available*

We are considering to publish the EUR dataset in the UK Data Archive or Zenodo.
Specify what methods or software tools are needed to access the data?

- *Name the required methods or software tools*

Quantitative data in SPSS format (.sav) can be accessed with IBM SPSS (https://www.ibm.com/analytics/us/en/technology/spss/), and open source data analysis software such as R (https://www.r-project.org/). Both software allow for exporting the data set in a number of other formats, including Microsoft Excel format (.xls, xlsx) and CSV.
Qualitative data in Microsoft Word format (.doc, .docx) can be accessed with Microsoft Word and open source text editing software such as Apache OpenOffice (https://www.openoffice.org/).

- *Is the software pre-existing or developed as an output of CARESSES*

All the applications mentioned above are pre-existing and independent from CARESSES.

- *Is documentation about the software available to access the data included?*

Most of the applications mentioned above come with rich documentation and support functionalities (see for example R https://cran.r-project.org/manuals.html).

- *Is it possible to include the relevant software (e.g. in open source code)?*

The applications mentioned above which are not open source (e.g. IBM SPSS, Microsoft Word) provide free trials.

2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

Dataset 1: Cultural Knowledge Base (CKB)

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability
Ontologies themselves are a tool for interoperability. Within CARESSES, the CKB ontology constitutes the vocabulary for the culturally competent assistive robot to be developed in the course of the project and facilitates the use and interaction among all

software tools developed within the project. In its construction, whenever possible, we will adopt terms and definitions which are standard in the field or culture they refer to.

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?
When and however possible, we will refer to standard vocabularies.

Dataset 2: Interaction Logs (IL)

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability
Zenodo complies with several standard metadata format such as MARCXML, Dublin Core and DataCite Metadata Schema (http://about.zenodo.org/policies/). Moreover, the CSV format is among the most readable formats for information storage, supported by the vast majority of software for numerical and data analysis.

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?)
We will provide mapping to the CKB ontology, as well as other existing vocabularies, whenever possible.

Dataset 3: End-Users Responses (EUR)

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability
Zenodo complies with several standard metadata format such as MARCXML, Dublin Core and DataCite Metadata Schema (http://about.zenodo.org/policies/).
Quantitative data in the EUR dataset will comply with the data vocabulary of the tools they refer to, thus ensuring exchange and re-use by any researcher making use of the same or compatible tools. We will try to adhere, as far as our application permits it, to the European Language Social Science Thesaurus (ELSST - https://elsst.ukdataservice.ac.uk/elsst-guide/elsst-structure).

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?
We will provide mapping to the ELSST thesaurus, the CKB ontology, as well as other existing vocabularies, whenever possible.

2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**
- **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

Dataset 1: Cultural Knowledge Base (CKB)

Specify how the data will be licensed to permit the widest reuse possible

- *Who owns the data?*

The matter of the ownership of data produced within the project is discussed in the Coordination Agreement among partners. This matter will be handled under the supervision of the Exploitation, Dissemination and IPR board.

- *How will the data be licensed for reuse?*

Licensing terms will be defined by the CARESSES partners and in accordance with the restrictions, if any, of any third-party data used in the CKB ontology.

- *If you are using third-party data, how do the permissions you have been granted affect licensing?*

A number of ontologies (such as the Time Ontology from W3C) grant “permission to copy, and distribute their contents in any medium for any purpose and without fee or royalty”. We will keep the licensing terms of any third-party ontology we will use in the CKB ontology into account in the definition of the licensing terms of the CKB ontology itself.

- *Will data sharing be postponed / restricted e.g. to seek patents?*

Probably not. However, it will most likely be postponed to comply with publication regulations. This matter will be handled under the supervision of the Exploitation, Dissemination and IPR board.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed (no later than publication of the main findings and should be in-line with established best practice in the field)
According to the CARESSES Work plan, the CKB ontology will be ready for publication approx. from month 25 (third year of the project). The CKB ontology will be officially publicly released upon the publication of related articles.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project?

- *Who may be interested in using your data?*

As previously stated, we envision the CKB ontology to be especially useful for: 1) researchers in the field of Robotics, who may use it as a guideline and reference for the development of robots able to interact with people while keeping cultural information into account; 2) companies producing robots and other devices for personal assistance, who may use it as a source of validated information for a number of cultures (specifically, the ones to be considered during the testing phase of CARESSES), allowing for culture-aware human-robot interaction; 3) researchers and practitioners in the field of Transcultural Nursing, who may use it as a validated and publicly available ontology for the description of concepts related to cultural competence and the detailing of a number of cultures.

- *What are the further intended or foreseeable research uses for the data?*

See above

- *If the re-use of some data is restricted, explain why*

At the moment, we do not foresee any restriction on the re-use of the CKB ontology.

Describe data quality assurance processes
"Data quality" can be defined in terms of syntactic, semantic and pragmatic quality (see ISO 8000-8:2015).
A number of ontology editors, such as Protégé, provide tools for automatically detecting inconsistencies in the ontology and checking its validity. Moreover, there exist publicly available tools, such as Oops! (http://oops.linkeddata.es/) which automatically check for anomalies, errors and lack of metadata for documentation. As an example, the full catalogue of pitfalls detected by Oops! Is available at http://oops.linkeddata.es/catalogue.jsp. The pragmatic quality of the CKB ontology (i.e., whether it fits for its intended use) will be checked during its creation by the experts involved in the CARESSES project, and experimentally evaluated in the testing phase of the project.

Specify the length of time for which the data will remain re-usable
Forever.

Dataset 2: Interaction Logs (IL)

Specify how the data will be licensed to permit the widest reuse possible

- *Who owns the data?*

The matter of the ownership of data produced within the project is discussed in the Coordination Agreement among partners. This matter will be handled under the supervision of the Exploitation, Dissemination and IPR board.

- *How will the data be licensed for reuse?*

Licensing terms will be defined by the CARESSES partners.

- *If you are using third-party data, how do the permissions you have been granted affect licensing?*

We do not foresee the use of any third-party data in the IL dataset.

- *Will data sharing be postponed / restricted e.g. to seek patents?*

Probably not. However, it will most likely be postponed to comply with publication regulations. This matter will be handled under the supervision of the Exploitation, Dissemination and IPR board.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed (no later than publication of the main findings and should be in-line with established best practice in the field)
According to the CARESSES Work plan, Interaction Logs are collected in two separate stages of the project: first in the course of Task 5.6 (m23 – m27) and then in the course of Task 6.3 (m28-m33) and Task 6.4 (m28-m33). As such, the first portion of the IL data set will be ready for publication approx. from month 27, while the second portion of the IL data set will be ready for publication approx. from month 37(end of the project) and it will be officially publicly released upon the publication of related articles.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project?

- *Who may be interested in using your data?*

As previously stated, we envision the IL data set to be useful for academic and industrial researchers aiming at defining guidelines,

best practices and standards in the field of Human-Robot Interaction (e.g., identifying which robot actions are frequently requested by people, identifying recurrent sequences of robot actions – human actions that a robot could rely on to exhibit predictive behaviours). Portions of the dataset may also be used by roboticists for the development and testing of specific robotic applications (e.g., the IL dataset can be used to train and test algorithms for learning the habits/routines of a person from the analysis of recurring events).

- *What are the further intended or foreseeable research uses for the data?*

See above

- *If the re-use of some data is restricted, explain why*

At the moment, we do not foresee any restriction on the re-use of the IL data set.

Describe data quality assurance processes
"Data quality" can be defined in terms of syntactic, semantic and pragmatic quality (see ISO 8000-8:2015) or, in other words, in terms of completeness, validity, accuracy, consistency.
Data completeness indicates whether all the data necessary to meet the current (and possibly future) information demand are available. By design, the IL data set fulfills the requirements of the culture-aware robot developed in the CARESSES project, thus ensuring that it contains sufficient information for an assistive robot to have meaningful interactions with a person. Data validity will be assessed in WP2, WP3 and WP4, as part of the development process of the software modules producing the messages to be stored in the IL data set. Data accuracy and consistency refer to whether the values stored are correct or not. Since the data to be stored in the IL data set are used by the culturally competent robot to tune its behavior towards the assisted person, one of the goals of the project is to maximize their reliability. To allow for a quantitative assessment of the accuracy of the data in the IL data set, we will consider providing, together with the portion of the IL data set acquired in Task 5.6 in lab conditions, supporting material providing the ground truth of the stored data.

Specify the length of time for which the data will remain re-usable
Forever.

Dataset 3: End-Users Responses (EUR)

Specify how the data will be licensed to permit the widest reuse possible

- *Who owns the data?*

The matter of the ownership of data produced within the project is discussed in the Coordination Agreement among partners. This matter will be handled under the supervision of the Exploitation, Dissemination and IPR board.

- *How will the data be licensed for reuse?*

Licensing terms will be defined by the CARESSES partners.

- *If you are using third-party data, how do the permissions you have been granted affect licensing?*

We do not foresee the use of any third-party data in the EUR dataset.

- *Will data sharing be postponed / restricted e.g. to seek patents?*

Probably not. However, it will most likely be postponed to comply with publication regulations. This matter will be handled under the supervision of the Exploitation, Dissemination and IPR board.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed (no later than publication of the main findings and should be in-line with established best practice in the field)
According to the CARESSES Work plan, End-Users Responses are defined, structured and collected in the course of Tasks 6.1, 6.2 and 6.3, which span months 19 to 33 of the project. Quantitative data are then post-processed and analysed in the course of Tasks 7.1 and 7.3, which span months 27 to 37 of the project, while qualitative data are post-processed and analysed in the course of Tasks 7.2 and 7.3, which span months 29 to 37. The EUR data set will therefore be ready for publication approx. from month 37(end of the project) and it will be officially publicly released upon the publication of related articles.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project?

- *Who may be interested in using your data?*

We envisage that the EUR data will be useful for academics interested in conducting secondary analysis. This could include academics interested in the acceptability and clinical or cost-effectiveness impact of culturally aware robots, but also those interested in our baseline characteristics and outcome measurement data.

- *What are the further intended or foreseeable research uses for the data?*

See above

- *If the re-use of some data is restricted, explain why*

At the moment, we do not foresee any restriction on the re-use of the EUR data set.

Describe data quality assurance processes

“Data quality” can be defined in terms of syntactic, semantic and pragmatic quality (see ISO 8000-8:2015) or, in other words, in terms of completeness, validity, accuracy, consistency.

We will strive for data completeness by constructing methodological protocols and tools that are user-friendly and sensitive. For the quantitative data, to increase the likelihood of validity and accuracy, we will employ existing widely used, previously validated data collection instruments such as the SF-36 and ZBI. For all of the quantitative tools we employ, we shall conduct a series of Cronbach’s Alpha-coefficient tests in SPSS. This will also help with establishing internal consistency. Further, we shall conduct Cohen’s kappa tests to establish the degree of inter-rater consistency between the researchers collecting data. To help boost the likelihood of consistency being achieved, the research team will be trained to follow the same strict protocols throughout. For qualitative data, to help boost the trustworthiness of our analysis we should engage in respondent validation exercises with our participants. Our interview schedules and data collection processes will be sensitive and planned so that they are likely to be complete and effective.

Specify the length of time for which the data will remain re-usable

Forever.

3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

Dataset 1: Cultural Knowledge Base (CKB)

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs (œsts related to open access to research data are eligible as part of the Horizon 2020 gran)

The costs of making the CKB ontology FAIR are included in Task 2.6.

Describe costs and potential value of long term preservation

Once the CKB ontology is publicly available, the only foreseeable cost for its preservation is the cost of the repository hosting service where it is located.

Dataset 2: Interaction Logs (IL)

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs (œsts related to open access to research data are eligible as part of the Horizon 2020 gran)

The costs of making the IL data set FAIR are included in Task 5.6 (for the first portion of the data set) and in Task 7.3 (for the second portion of the data set). The cost and effort of making the second portion of the data set FAIR is expected to be significantly lower than that of the first portion of the data set.

Describe costs and potential value of long term preservation

Once the IL data set is publicly available, the only foreseeable cost for its preservation is the (eventual) cost of the repository hosting service where it is located.

Dataset 3: End-Users Responses (EUR)

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs (œsts related to open access to research data are eligible as part of the Horizon 2020 grant)

SPSS and NVivo 11 licensing costs may be applicable. Otherwise we do not envisage any additional costs associated with making our data FAIR.

Describe costs and potential value of long term preservation

Once the data set is publicly available, the only foreseeable cost for its preservation is the (eventual) cost of the repository hosting service where it is located.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

Dataset 1: Cultural Knowledge Base (CKB)

Specify if the data should be safely stored in certified repositories for long term preservation and curation.)
We are considering applying for the inclusion of the CKB ontology in well known collections of Ontologies (we will apply for its insertion in popular ontology libraries and search engines, such as the Protégé Ontology Library, LOV and Google) to make it publicly available to a large audience. We will host the CKB ontology on a repository which provides adequate guarantees in terms of data persistence, security and accessibility.

Is your data sensitive (e.g. detailed personal data, politically sensitive information or trade secrets)?(YES/NO)
NO

Dataset 2: Interaction logs (IL)

Specify if the data should be safely stored in certified repositories for long term preservation and curation.
We are considering archiving the IL data set with the Zenodo data archiving tool to ensure long term preservation and curation.

Is your data sensitive (e.g. detailed personal data, politically sensitive information or trade secrets)? (YES/NO)
NO

Dataset 3: End-Users Responses (EUR)

Specify if the data should be safely stored in certified repositories for long term preservation and curation.
We are considering archiving the EUR data set with the Zenodo data archiving tool to ensure long term preservation and curation.

Is your data sensitive (e.g. detailed personal data, politically sensitive information or trade secrets)? (YES/NO)
NO

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Dataset 1: Cultural Knowledge Base (CKB)

Are the data acquired by carrying out research involving human participants?(YES/NO)
YES

If the answer is YES,

- *Specify the procedure established to gain consent for data preservation and sharing*

Human participants will be involved in the collection and validation of culture-specific information. Participants' responses will be merged and generalized, and no person-specific detail will be stored in the CKB ontology (which, by design, captures cultural information at a national/group level). Therefore, data not fall under the General Data Protection Regulation (EU) 2016/679. All of the data will be collected in an ethically appropriate manner and with Ethics Committee approval.

- *Specify how will sensitive data be handled to ensure it is stored and transferred securely*

No sensitive data will be stored in the CKB ontology.

- *Specify how will you protect the identity of participants, e.g. via anonymisation or using managed access procedures*

No personal data about the participants will be stored in the CKB ontology.

Are the data acquired by carrying out research involving human participants? (YES/NO)
YES

Dataset 2: Interaction logs (IL)

If the answer is YES,

- *Specify the procedure established to gain consent for data preservation and sharing*

Human participants will be involved in the collection of recordings of interactions between the culturally competent robot and a person.
By design, the IL data set does not contain any person-specific detail, since it only captures events and status information which are of relevance for the robot to plan and tune its behavior. Moreover, messages refer to participants only by an ID which ensures the protection of their identity both during the experiments and in the public IL data set. Therefore, data not fall under the General Data Protection Regulation (EU) 2016/679. All of the data will be collected in an ethically appropriate manner and with Ethics Committee approval.

- *Specify how will sensitive data be handled to ensure it is stored and transferred securely*

No sensitive data will be stored in the IL data set.

- *Specify how will you protect the identity of participants, e.g. via anonymisation or using managed access procedures*

No personal data about the participants will be stored in the IL data set. Moreover, participants will be exclusively identified by an ID.

Dataset 3: End-Users Responses (EUR)

Are the data acquired by carrying out research involving human participants?
YES
If the answer is YES,

- *Specify the procedure established to gain consent for data preservation and sharing*

Ethical approval from appropriate bodies will be defined before the experiments and participants will be asked to provide informed consent for their participation in the study. This will include consent for data preservation and sharing.

- *Specify how will sensitive data be handled to ensure it is stored and transferred securely*

The data will be collected following informed consent and will be pseudonymized, in compliance with the General Data Protection Regulation (EU) 2016/679.

- *Specify how will you protect the identity of participants, e.g. via anonymisation or using managed access procedures*

As above (via pseudonymized)

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

We do not consider other procedures for data management.