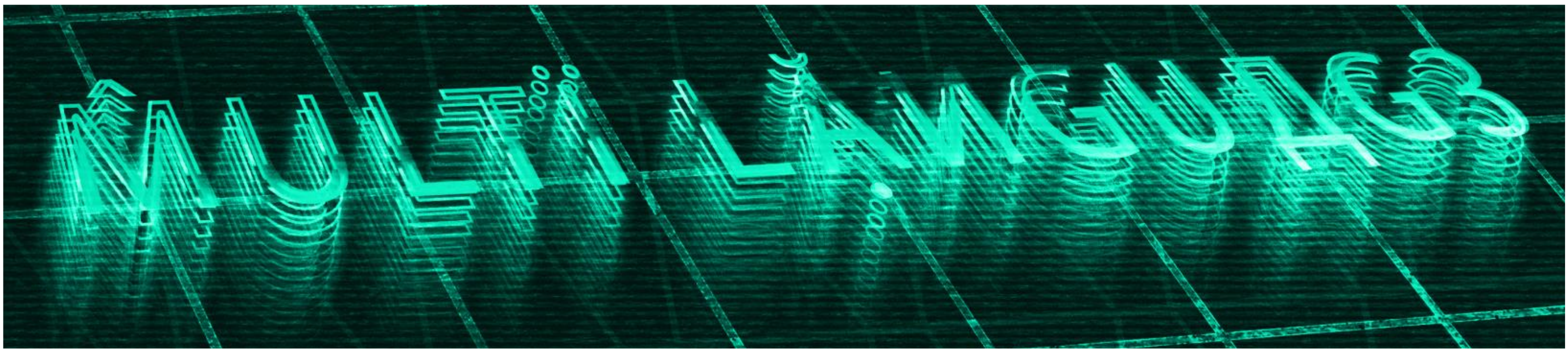

Multi-Language Messages Dataset

Dataset Card

Version 2.0

Last Updated: 23rd February 2024

Contents

How to use this document

Purpose

This document contains information about the Multi-Language Messages dataset, to help users understand what the dataset contains, its intended use cases, how it was developed, associated metadata, and information on how to use the data responsibly.

The Dataset Card consists of three main parts: an overview of the dataset, dataset structure, and information about the curation methodology. This includes information on privacy, ethics, and any known biases and limitations of the dataset.

This document is intended for use primarily by the data science community.

Navigation

Use the hyperlinked tabs along the left-hand side to navigate through the document, and between pages and sections.

Dataset Description
Dataset Structure
Curation Methodology
Licensing Information
Additional Information

Dataset description

23 billion text messages are sent each day. That works out to 270,000 messages each second. In fact, by the time you have finished reading this sentence over 2.35 million messages will have been sent!

Messaging content is extremely diverse. More than half of the world's population speak at least two languages, and it's common to mix between different languages. A convenient way when writing in multiple languages is to use romanisation, where words are converted from one language script to Roman, i.e., Latin, script. This may be for convenience, for example to enable use of a single keyboard, to create emphasis, or simply when writing 'stream of consciousness'.

For example: "Marhaba, how are you today?" can be simpler to write out than "مرحبا, how are you today?". This example uses transliteration, where the Arabic word "مرحبا" is romanised into English as "marhaba". The romanised word represents its Arabic pronunciation as closely as possible.

This **Multi-Language (ML) messages dataset** is made up of SMS-style conversations in English, Arabic, Farsi and Russian. Messages are either in the language's native script, the romanised equivalent, in English, or a mix of the above.

The dataset has 2,196 conversations, which contain a diverse range of topics and authors from a wide range of demographics. As they are SMS-style, the messages contain misspelt words, slang, and informal chat-language.

In total the dataset contains 32,533 messages, written across one of the following languages:

1. Arabic
    1. Gulf (2,853 messages)
    2. Levantine (4,312 messages)
    3. Egyptian (3,923 messages)
2. Farsi (10,095 messages)
3. Russian (10,641 messages)
4. English (709 messages)

Dataset Description
Dataset Structure
Curation Methodology
Licensing Information
Additional Information

The dataset contains labels for:

1. The language that each message is written in.
2. Whether the message is written in the native script, romanised or mixed script / language.
3. Word-level span labels, which specify the language label for each word within the messages.
4. The topic of the conversation.

This dataset contains text only**,** and is stored in a csv file.

Table 1: Common terminology used within this dataset card

Terminology	**Definition**	**Example**
Span	A pair (tuple) of integers representing the start and end index of a word. Each span is at the word level, meaning that one word corresponds to one span.	• "Hello, سال salám" has three words and will thus have three spans: ○ "Hello" is English and has the span (0,5) ○ "سال" is Native Farsi and has the span (7,10) ○ "salám" is romanised Farsi and has the span (11,16)
List of spans	A list of span labels	[(4,7), (8,12)]
Native script	This is a language that is written in its native script. The dataset has messages written using the Perso-Arabic, Cyrillic and Arabic scripts for Persian, Russian and Arabic words respectively.	The Farsi word: "سال"

Romanisation	This is a process of converting a language that uses a non-Latin script into Roman (Latin) text, typically based on the phonetics of the word.	The Farsi word: "سال" can be written phonetically using Latin letters as "saal"
Special	Special characters, such as Unicode emojis or text-based emojis.	🤓 or ":D"
Ambiguous	A span that could not be defined as Arabic, English, Farsi, Russian or English, even when considering possible spelling mistakes	"asnkjdnada" is ambiguous as it does not belong to any language.
AWS Ground Truth	This is an AWS service that allows its users to label data.	

Supported tasks

This dataset can be used to train:

- Language detection models (message and word level)
- Romanisation detection models (message and word level)
- Multi-language topic models

Dataset structure

Data format

The data is in the form of a csv that has the structure shown in table 2:

Table 2: Sample dataset rows

index	message	conversation_id	topic_id	romanisation_label	lang_label	english	romanised_arabic	native_arabic	romanised_russian	native_russian	romanised_farsi	native_farsi	special
0	Alykom Al Salam, Aywa Enty betheby El books we El Rewayat kteer.	6607940560	discuss_hobbies	mixed	en, ar-Eg	[(38, 43), (44, 46)]	[(0, 6), (7, 9), (10, 15), (17, 21), (22, 26), (27, 34), (35, 37), (47, 49), (50, 57), (58, 63)]	[]	[]	[]	[]	[]	[]
1	Aywa bgad ana aker haga can makes me happy hya el reading.	6607940560	discuss_hobbies	mixed	en, ar-Eg	[(24, 27), (28, 33), (34, 36), (37, 42), (50, 57)]	[(0, 4), (5, 9), (10, 13), (14, 18), (19, 23), (43, 46), (47, 49)]	[]	[]	[]	[]	[]	[]

Data fields within the dataset

Table 3: Columns of the dataset

Column name	Data type	Description
index	Integer	Unique identifier for each message.
message	String	A message from a conversation, typically ~6 words long.
conversation_id	Integer	ID of a conversation.
topic_id	String	ID of the topic of the conversation that the message is a part of, these IDs are listed in Table 5.
group_id	String	ID of the group that generated the conversation.
romanisation_label	String	A message-level label for whether the content is "romanised", "mixed", or "non_romanised". These labels are defined as follows: • **romanised**: a message where **all** words are romanised (e.g., the message is written fully in romanised Farsi). • **non_romanised**: a message where **all** words are written in a language's native script. Note, English words are considered non-romanised. These messages may contain mixed language content (i.e., English and native Farsi/Arabic/Russian, but will not contain any romanised Farsi/Arabic/Russian). • **mixed**: a message where the content is a mixture of romanised and non-romanised words. These messages may contain mixed language content (i.e., English and words written in romanised Farsi/Arabic/Russian) or single language content (i.e., romanised Farsi/Arabic/Russian and native Farsi/Arabic/Russian). This label should be used in combination with the lang_label (below) to identify whether content is mixed script and mixed language, or just mixed script.

Dataset Description
Dataset Structure
Curation Methodology
Licensing Information
Additional Information

lang_label	String	The language(s) a message is written in. Where there are two languages present in the message, these are recorded as a comma separated string e.g., en, ar-Le. Note that there will only ever be English present as the second language in a message. The language labels in the dataset are: • **ar-Gu**: Gulf Arabic • **ar-Le**: Levantine Arabic • **ar-Eg**: Egyptian Arabic • **ru**: Russian • **fa**: Farsi • **en**: English • **en, ar-Gu**: English and Gulf Arabic • **en, ar-Le**: English and Levantine Arabic • **en, ar-Eg**: English and Egyptian Arabic • **en, ru**: English and Russian • **en, fa**: English and Farsi
english	List of tuples	A list of spans that contain the start and end indexes of English words in the message.
romanised_arabic	List of tuples	A list of spans that contain the start and end indexes of romanised Arabic words in the message.
native_arabic	List of tuples	A list of spans that contain the start and end indexes of Native script Arabic words in the message.
romanised_russian	List of tuples	A list of spans that contain the start and end indexes of romanised Russian words in the message.
native_russian	List of tuples	A list of spans that contain the start and end indexes of Native script Russian words in the message.
romanised_farsi	List of tuples	A list of spans that contain the start and end indexes of romanised Farsi words in the message.
native_farsi	List of tuples	A list of spans that contain the start and end indexes of Native script Farsi words in the message.

Dataset Description
Dataset Structure
Curation Methodology
Licensing Information
Additional Information

special	List of tuples	A list of spans that contain the start and end indexes of Special Text, such as emojis, in the message.

Class counts in the dataset

The dataset has labels that denote the languages and romanisation present in each message, as well as topic labels at a conversation level. The tables below present the class counts for these labels.

Note that the combination of English and native text is classed as non-romanised, therefore non-romanised spans exist in mixed language labels (e.g., en, fa). To derive the sub-classes for messages that are a mix of languages and/or scripts, the language labels and romanisation labels must be used together. For example:

- **single language** and **single script** will have one language label listed (e.g., ar-Eg) and the romanisation label will be romanised (e.g., a message containing only words that are in romanised Egyptian Arabic).
- **single language** and a **mixed script** will have one language listed (e.g., ar-Eg) and the romanisation label will be mixed (e.g., a message containing words that are written in a mix of romanised and native script Egyptian Arabic).
- **mixed language** and **single script** messages will have two languages listed in the language label (e.g., en, ar-Eg) and the romanisation label will be non-romanised (e.g., a message containing words that are in English and native script Egyptian Arabic. English and native script Egyptian Arabic are both considered non-romanised).
- **mixed language** and **mixed script** messages will have two languages listed in the language label (e.g., en, ar-Eg) and the romanisation label will be mixed (e.g., a message containing words that are in English and romanised Egyptian Arabic. English text is classed as non-romanised).

Table 4: Languages and romanisation label message level counts

Language label	Romanisation label	Message count
Arabic Egyptian	non-romanised	149

	romanised	1,911
English and Arabic Egyptian	mixed	1,842
	non-romanised	21
Arabic Gulf	non-romanised	99
	romanised	1,186
English and Arabic Gulf	mixed	1,565
	non-romanised	3
Arabic Levantine	non-romanised	89
	romanised	2,134
English and Arabic Levantine	mixed	2,067
	non-romanised	22
English	non-romanised	709
Farsi	mixed	38
	non-romanised	1,109
	romanised	4,396
English and Farsi	mixed	4,450
	non-romanised	102
Russian	non-romanised	778
	romanised	4,595
English and Russian	mixed	5,233

	non-romanised	35

Table 5: Counts of messages for the conversation topic labels

Topic description	**Topic_id**	**% Overall Dataset**	**Total count**
Unspecified	random	45.9%	14,931
Organising to attend a fictional event in the future e.g, cinema, festival, meal	organise_event	8.4%	2,719
Complaining about your job to a friend	complain_job	7.1%	2,319
Teaching someone about something you are interested in	teach_general	7.4%	2,422
Talking about what they did last weekend (e.g., visit an art gallery / museum)	discuss_weekend	7.2%	2,335
Discussing the weather over the last week	discuss_weather	8.7%	2,839
Debate around something you care about e.g, football team, favourite sport / book / film	debate_general	7.9%	2,576
Discussing your hobbies together	discuss_hobbies	7.4%	2,392
Total		**100%**	**32,533**

Curation methodology

Curation rationale

There is a notable absence of "gold-standard" labelled romanised corpuses for the languages in this dataset. Not only is data sparse in this area, but labelled datasets of informal chat-style text are particularly lacking, despite it being a common and everyday use of language.

Most romanised datasets that do exist are taken from public information websites, e.g., Wikipedia, rather than informal content or social media contexts. Text content written in SMS-style is very different to full prose, which leads to solutions that have been trained on full sentences of text underperforming when presented with more 'noisy' text seen in informal conversational-style text.

This dataset is the first of its kind in providing a large-scale labelled SMS-style dataset, which is a combination of English, Arabic, Russian and Farsi and contains both native script as well as romanised content for those languages.

Curation approach

The data was curated in two stages:

1. Conversation generation

The conversations were curated through crowdsourcing, with each participant simulating a SMS-style conversation, as if they were texting a friend or relative. Each participant generated conversations in a combination of one language (Arabic (Gulf, Levantine, Egyptian), Russian, or Farsi) and English, while switching between romanised and non-romanised forms of the non-English language and English (e.g., mixing words in Arabic and English within a message). The messages may contain emojis, spelling mistakes, acronyms, and other non-standard text features present in SMS-style messaging. The dataset contains no non-text content (GIFs, photos etc), and also does not contain any personally identifiable information.

To facilitate natural, informal and varied conversations, crowd members were given a conversation prompt topic for 58% of the messages (see topics listed in

Dataset Description
Dataset Structure
Curation Methodology
Licensing Information
Additional Information

Table 5, Class Counts Section above). These topics were intended starting points and from that point, conversations could expand to cover whichever topics the crowd members wished.

Finally, to create greater variation in the content generated, the crowd comprised of participants from different regions, genders and age groups. Overall, there were 564 unique authors and 3,999 unique conversations generated. The average message length is 7 words, and the average conversation length is 20 messages.

Table 6: Participant Demographic Information

Language / Dialect	Regional distribution of authors	# of authors
Arabic Gulf	• UAE 73% • Saudi Arabia 18% • Qatar 1% • Others (Algeria, Bahrain, Egypt, Jordan, Morocco, Oman) 8%	48
Arabic Egyptian	• Egypt 85% • United Kingdom 13% • United States 3%	34
Arabic Levantine	• Jordan 71% • Lebanon 29%	44
Farsi	• UAE 100%	103
Russian	• Kazakhstan 55% • United Kingdom 20% • Ukraine 14% • Other (Armenia, Georgia, Portugal, Turkey) 11%	113
English	• Not provided due to privacy	1

2. Data annotation

After the conversation took place, the messages were retrospectively labelled by another native speaker of that language and the word-level language spans added. The method to annotate the dataset was as follows:

Message level labels:

- **Topic (topic_id)** - was pre-specified before the conversation was generated. This topic will be the same across all messages within a conversation.
- **Message Language (lang_label)** - human-specified by a native speaker of the language, after the conversation was generated.
- **Romanisation Label (romanisation_label)** – human-specified by a native speaker of the language, after the conversation was generated.

Word level labels:

The spans were curated in a pipeline that was different depending on whether the message was labelled as 'mixed' or written in a single language. All messages were split using white space to distinguish words within the dataset, then for:

1. **Single language messages** each word span was labelled with the message-level language (e.g., for a Farsi message, all of the words were labelled as 'fa')
2. **Mixed messages** the word spans within the messages were labelled using the data labelling pipeline detailed below.

Data Labelling Pipeline Steps:

1. An automated process was used to label common English words (using a standard corpus of English words from the PyEnchant library https://pypi.org/project/pyenchant/). Any known 'overlap' words were identified (e.g., 'man' has meaning in both Farsi and English).
2. Two humans, who were a mix of native English speakers and individuals fluent in the non-English languages, reviewed the automated span-labels. These reviewers assessed all sentences to determine whether the word labels were correct. If both reviewers agreed on the accuracy of the automated labels, then the span labelling was approved.

3. Any messages where the automated labels were identified as incorrect by at least one reviewer were sent to linguists to be relabelled. AWS Ground Truth was used to label each message (this varied between the languages but on average was ~10% of the dataset). Each message was re-labelled by two individual linguists, and their annotations were compared. If their annotations matched, the span labelling was approved.
4. Any remaining mismatching annotations were then reviewed by a Linguist and any annotation errors corrected.

Data privacy and ethics

All participants gave informed consent to participate in this data collection, with the knowledge that the conversational data could be released publicly and used for training and developing data science models.

All participants were assigned an ID number and no demographic identity categories (including but not limited to age, gender or country) are shared alongside this dataset, meaning that is not possible to trace back to any individual participant.

All participants were paid the living wage for their region for participation in the data curation.

During data collection, participants were instructed that they should **not** share any Personally Identifiable Information. This was validated through quality control checks on the conversational data. These are some examples of PII shown to participants for them to avoid disclosing:

- Full names
- Credit/bank information
- Identity documentation (ID Numbers for example Social Security, account numbers or other citizenship numbers)
- Personal addresses
- Contact information
- Medical information
- Information about education (such as schools attended) or profession (such qualifications, place of employment)

- Information about racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union data
- Data concerning sexual orientation
- Data concerning criminal allegations, proceedings or convictions.

Participants were also instructed not to talk about anything deemed offensive. This was validated through quality control checks on the conversational data, with examples given being:

- Defamatory content
- Violence or inciting violence
- Hate speech
- Indecent images

PII disclaimer
Disclaimer: We have taken best endeavours to remove all Personally Identifiable Information (PII) from this dataset. If you identify any conversations with content that could be deemed as PII please contact challengeaiuk@deloitte.co.uk.

Biases

The dataset was purposefully designed to contain much more romanised text in the messages than would typically be found in real world data. This was done to increase the proportion of romanised content, as there was a specific data gap for this type of content. This may mean that conversations are not as naturally representative of typical conversations in terms of romanisation / mixing languages, however they should still flow in a natural way.

There is also a small regional bias within the data, with speakers of the non-English languages coming from a small number of regions that speak the language (see Table 6 in the Curation Method section above). This may present as a lack of local colloquial content (e.g., idioms or common sayings) from the under-represented regions, although given the number of unique contributors this is not expected to present as a large bias that could impact model performance.

Bias notification
Disclaimer: We have taken best endeavours to design out any other biases from this dataset. If you identify any other biases please contact challengeaiuk@deloitte.co.uk.

Other known limitations

Some minor errors may persist due to the human error of labelling, this error may be present in ~1-2% of the dataset. Possible errors include small words or spans being misclassified with a language, particularly punctuation. For messages that may contain an error, an average of 1 out of every 6 spans is estimated to be impacted.

Some span labels may appear to contain errors, but are actually correct due to typos in the human-written messages (which is representative of a real life scenario). Effort has been taken to correct as many of these errors as is possible, however some will remain as they are unique edge cases.

Licensing Information

PLEASE READ THE TERMS BELOW CAREFULLY BEFORE USING THE DATASET

What's in these terms?

This data licence ("**Data Licence**") applies to

1. any and all Data (as defined below) that you download from the Site (as defined below) and
2. any content or materials you create that are based on, or contain (in whole or part), any such Data.

We own, or license the use of, the Intellectual Property Rights (as defined below) in the Data and agree to provide the Data to you on the terms set out in this Data Licence.

1. Who we are and how to contact us

1.1 The challengeai.com website (the "**Site**") is operated by the Ministry of Defence. Our main address is Whitehall, London SW1A 2HB.

1.2 To contact us, please use the Get in contact button at the bottom of this website.

2. By downloading the Data you accept these terms

2.1 By downloading Data, you confirm (on behalf of your organisation) that you accept the terms of this Data Licence and that you agree to comply with them. You also confirm that you have the authority from your organisation to accept this Data Licence.

2.2 If you do not agree to these terms, you must not download any Data.

2.3 We recommend that you print a copy of these terms for future reference.

3. Interpretation

3.1 The definitions and rules of interpretation in this clause apply in this Data Licence.

- **Challenge**: any Challenge listed on the Site in the 'Challenges' section, which can be found here.
- **Computer System**: any information technology hardware, software and/or system owned, controlled, or operated by you/your organisation to which Data is received in accordance with this Agreement.
- **Data**: the datasets and associated information, documentation and materials which relate to any Challenge and are made available for download by users of the Site in order to participate in any such Challenge.
- **Derived Data**: any Data (wholly or in part) which is Manipulated to such a degree that it: (a) cannot be identified as originating or deriving directly from the Data and cannot be reverse-engineered such that any part of the underlying Data can be identified; and (b) is not capable of use substantially as a substitute for the Data.
- **Distribute**: to make Data accessible including the provision of access through a database or other application populated with the Data, resell, sub-license, transfer, disseminate, or otherwise disclose or provide a copy of the Data by any means in any format whether electronic or hard copy.
- **Good Industry Practice**: in relation to any undertaking and any circumstances, the exercise of skill, diligence, prudence, foresight and judgement and the making of any expenditure that would reasonably be expected from a skilled person engaged in the same type of undertaking under the same or similar circumstances.
- **Intellectual Property Rights**: all patents, rights to inventions, copyright and related rights, trade marks, service marks, trade, business and domain names, rights in trade dress or get-up, rights in goodwill or to sue for passing off, unfair competition rights, rights in designs, rights in computer software, database rights and other rights in data, moral rights, rights in confidential information (including know-how and trade secrets) and any other intellectual property rights, in each case whether registered or unregistered and including all applications for and renewals or extensions of such rights, and all similar or equivalent rights or forms of protection in any part of the world.
- **Licence**: the licence granted in clause 6.
- **Manipulate**: to combine or aggregate the Data (wholly or in part) with other data or information, or to adapt or change the Data (wholly or in part).
- **Manipulated Data**: any Data which has been Manipulated.

- **Permitted Purpose**: the purpose in respect of which you are permitted to use the Data, as set out in clause 6.1.

3.2 A **person** includes a natural person, corporate or unincorporated body (whether or not having separate legal personality).

3.3 A reference to **writing** or **written** includes email, but not fax.

3.4 Any words following the terms **including**, **include**, **in particular** or **for example** or any similar phrase shall be construed as illustrative and shall not limit the generality of the related general words.

4. There are other terms that may apply to you

4.1 This Data Licence applies to your use of Data, however the following additional terms (Additional Terms), also apply to your use of our Site generally:

- Our Website Terms and Conditions (Website Terms) that apply to your use of our Site.
- Our Privacy Policy. See further under How we may use your personal information.
- Our Acceptable Use Policy, which sets out the permitted uses and prohibited uses of our Site. When using our Site, you must comply with this Acceptable Use Policy.
- Our Challenge Terms on each relevant Challenge Rules page, which sets out the rules and requirements of participation in any Challenge.

5. We may make changes to this Data Licence

5.1 We may amend this Data Licence from time to time. Every time you wish to download Data from this Site, please check these terms to ensure you understand the terms that apply at that time. These terms were most recently updated on 6th February 2024.

6. Licence

6.1 Subject always to your compliance with the terms of this Data Licence, we grant to your organisation a non-exclusive, non-transferable, worldwide, royalty-free, perpetual but revocable licence to:

(a) access, view, use, Manipulate the Data and create Derived Data; and

(b) store copies of the Data and Manipulated Data on your Computer Systems,

solely for: (i) the purposes of your participation in any Challenge; (ii) the purpose of demonstrating a solution or capability which does not form part of a Challenge ; and/or (iii) your organisation's own internal business purposes, the limitations on which are set out below ("**Permitted Purpose**"). For the avoidance of doubt 'internal business purposes' means the usage of the Data solely by you, only on Computer Systems controlled by you/your organisation, in relation to the internal business operations of your organisation (including internal research and development, and internal training). Please note that this is not a personal licence granted to you as an individual.

6.2 Except as expressly provided in this Agreement, you shall (personally and on behalf of your organisation):

(a) limit access to the Data to those individuals who need to access it in order to carry out the Permitted Purpose;

(b) only make copies of the Data to the extent reasonably necessary for the Permitted Purpose, including where necessary, disaster recovery and testing;

(c) not use the Data in your personal capacity, outside the performance of your duties for your organisation.

(d) not use the Data for any purpose contrary to any law or regulation or any regulatory code, guidance or request;

(e) not extract, reutilise, use, exploit, Distribute, modify, copy, publish or store the Data (in whole or in part), for any purpose not expressly permitted by this Agreement. In particular you will not Distribute any of the Data to the public, third-parties and/or any person outside of your company or organisation;

(f) not decompile, reverse engineer or create derivative works from the Data for purposes other than the Permitted Purpose;

(g) not use the Data (wholly or in part) in your products or services that are made available to the public;

(h) not use the Data in a manner that allows or facilitates any Data to fall into the public domain;

(i) you shall not, nor directly or indirectly assist any other person to, do or omit to do anything to diminish our rights in the Data.

7. Data and Materials

7.1 We shall supply the Data on the Site.

7.2 You acknowledge that the Data is confidential, and you shall not disclose it, in whole or in part, to any third party, except as expressly permitted by this Data Licence. This obligation shall continue to apply after termination of this Data Licence.

7.3 At any time, we may, with as much prior notice to you as is reasonably practicable, change: (a) the content, format or nature of the Data; and (b) the means of access to the Data.

8. Security and passwords

8.1 You shall ensure that Data is kept securely and shall implement security practices and systems in accordance with Good Industry Practice to prevent, and take prompt and proper remedial action against, unauthorised access, copying, modification, storage, reproduction, display or Distribution of the Data.

8.2 If you become aware of any misuse of any Data, or any security breach (i.e. any incident that results in unauthorised access to computer data, applications/software, networks or devices) in connection with this Data Licence that could compromise the security or integrity of the Data or otherwise adversely affect us: (a) you shall, promptly notify us and fully co-operate with us to remedy the issue as soon as reasonably practicable; and (b) you agree to co-operate with our reasonable security investigations.

9. Export

You shall not export, directly or indirectly, any of the Data (or any products, documents or materials, including software, incorporating any such Data) to any location (whether physical or digital) that is: not within your direct control.

10. Intellectual Property Rights ownership

10.1 You acknowledge that we are the owner or the licensee of all Intellectual Property Rights, in and to the Data.

10.2 Subject to our rights in the Data and the terms of this Data Licence (which shall at all times prevail), your organisation shall own the Intellectual Property Rights to the Derived Data. Your use of the Derived Data shall at all times be subject to the Licence.

10.3 We draw your attention again to the Website Terms and Conditions and Challenge Terms which contain obligations that you have agreed to in the event that your entry is a Winning Challenge Response (as defined in the Challenge Terms). In particular, you have agreed that your organisation will automatically grant a 12-month licence to us of the Intellectual Property Rights in any Winning Challenge Response submitted by you (including all Derived Data and underlying documents). For further details see the Website Terms and Conditions and the Challenge Terms.

11. Warranties

11.1 We warrant that we have the right to license the receipt and use of Data.

11.2 Except as expressly stated in this Data Licence, all warranties, conditions and terms, whether express or implied by statute, common law or otherwise are hereby excluded to the fullest extent permitted by law.

11.3 We will take all reasonable steps, in accordance with Good Industry Practice not to introduce any vulnerability or virus into your Computer Systems, whether via the Site or Data or otherwise.

11.4 Without limiting the effect of this clause 11, we do not warrant that: (a) the supply of the Data will be free from interruption; (b) the Data will be suitable for your Computer Systems; or (c) the Data is accurate, complete, reliable, secure, useful, or fit for purpose.

12. Limitation of liability

12.1 Neither party excludes or limits liability to the other party for: (a) fraud or fraudulent misrepresentation; or (b) any matter in respect of which it would be unlawful for the parties to exclude liability.

12.2 Subject to clause 12.1, we shall not in any circumstances be liable whether in contract, tort (including for negligence and breach of statutory duty howsoever arising), misrepresentation (whether innocent or negligent), restitution or otherwise, for any of the following losses suffered by you as a result of your use of the Data: (a) any loss (whether direct or indirect) of profits, business, business opportunities, revenue, turnover, reputation or goodwill; (b) any loss or corruption (whether direct or indirect) of data or information; (c) loss (whether direct or indirect) of anticipated savings or wasted expenditure (including management time); or (d) any loss or liability (whether direct or indirect) under or in relation to any other contract.

13. Termination

13.1 Without prejudice to any rights that have accrued under this Data Licence or any of our rights or remedies, we may terminate this Data Licence with immediate effect by giving written notice to you if you commit a material breach of any term and fail to remedy that breach within a period of [30] days after being notified in writing to do so.

13.2 Any provision of this Data Licence that expressly or by implication is intended to come into or continue in force on or after termination shall remain in full force and effect.

13.3 On any termination of this Data Licence for any reason, you shall as soon as reasonably practicable, delete or destroy (as directed in writing by us) all copies (in whole or in part) of the Data and Manipulated Data whether stored in hard copy or electronically on your Computer Systems or otherwise. You will also delete the Data and Manipulated Data from any database, software program, documentation and other materials which contain (in whole or in part) copies of the Data and/or Manipulated Data.

14. Announcements

You shall not make, or permit any person to make, any public announcement concerning this Data Licence, any Challenges or Data without our prior written consent, except as required by law, any governmental or regulatory authority (including any relevant securities exchange), any court or other authority of competent jurisdiction.

15. Assignment

This Data Licence is personal to you, and you shall not assign, transfer, mortgage, charge, sub-contract, sub-license (except as permitted above) declare a trust of or deal in any other manner with any of your rights and obligations under this Data Licence without our prior written consent.

16. Waiver

16.1 A waiver of any right or remedy is only effective if given in writing and shall not be deemed a waiver of any subsequent right or remedy.

16.2 A delay or failure to exercise, or the single or partial exercise of, any right or remedy shall not waive that or any other right or remedy, nor shall it prevent or restrict the further exercise of that or any other right or remedy.

17. Remedies

Except as expressly provided in this Data Licence, the rights and remedies provided herein are in addition to, and not exclusive of, any rights or remedies provided by law.

18. Inadequacy of damages

Without prejudice to any other rights or remedies that we may have, you acknowledge and agree that damages alone would not be an adequate remedy for any breach by you of the terms of this Data Licence. Accordingly, we shall be entitled to seek the remedies of injunction, specific performance or other equitable relief for any threatened or actual breach of the terms of this Data Licence.

19. Severance

If any provision or part-provision of this Data Licence is or becomes invalid, illegal or unenforceable, it shall be deemed deleted, but that shall not affect the validity and enforceability of the rest of this Data Licence.

20. No partnership or agency

Dataset Description
Dataset Structure
Curation Methodology
Licensing Information
Additional Information

20.1 Nothing in this Data Licence is intended to, or shall be deemed to, establish any partnership or joint venture between any of the parties, constitute any party the agent of another party, or authorise any party to make or enter into any commitments for or on behalf of any other party.

20.2 You confirm that you are acting on your own behalf and not for the benefit of any other person.

21. Third-party rights

This Data Licence does not give rise to any rights under the Contracts (Rights of Third Parties) Act 1999 to enforce any term of this Data Licence.

22. Governing law

This Data Licence and any dispute or claim arising out of or in connection with it or its subject matter or formation (including non-contractual disputes or claims) shall be governed by and construed in accordance with the law of England and Wales.

23. Jurisdiction

Each party irrevocably agrees that the courts of England and Wales shall have exclusive jurisdiction to settle any dispute or claim arising out of or in connection with this Data Licence or its subject matter or formation (including non-contractual disputes or claims).

Additional information

A. Version History

Version 2: Addition of Farsi and Arabic (Egyptian) messages

Version 2 contains the addition of Farsi and Arabic (Egyptian) messages.