ReproHum #1018-09: Reproducing Human Evaluations of Redundancy Errors in Data-To-Text Systems

Filip Klubička[1], **John D. Kelleher**[2]
ADAPT Centre
Technological University Dublin[1], Trinity College Dublin[2]
filip.klubicka@adaptcentre.ie, john.kelleher@tcd.ie

Abstract

This paper describes a reproduction of a human evaluation study evaluating redundancies generated in automatically generated text from a data-to-text system. While the scope of the original study is broader, a human evaluation—a manual error analysis—is included as part of the system evaluation. We attempt a reproduction of this human evaluation, however while the authors annotate multiple properties of the generated text, we focus exclusively on a single quality criterion, that of redundancy. In focusing our study on a single minimal reproducible experimental unit, with the experiment being fairly straightforward and all data made available by the authors, we encountered no challenges with our reproduction and were able to reproduce the trend found in the original experiment. However, while still confirming the general trend, we found that both our annotators identified twice as many errors in the dataset than the original authors.

Keywords: human evaluation, reproduction, redundancy, data-to-text

1. Introduction

This report presents a reproduction of a human evaluation originally conducted and presented in the paper *Neural Pipeline for Zero-Shot Data-to-Text Generation* (Kasner and Dusek, 2022). The authors present an alternative approach for zero-shot data-to-text generation where they generate English text by transforming single-item descriptions with a sequence of modules trained on general-domain text-based operations: ordering, aggregation, and paragraph compression. They train pretrained language models for performing these operations on a synthetic corpus and show that their approach enables data-to-text generation from RDF semantic triples in zero-shot settings, which produce more semantically consistent output by avoiding noisy human-written references.

While the scope of their original study is much broader, a human error annotation is included as part of their system evaluation, described in Section 7.2 of their paper with results summarised in Table 5. In this evaluation step the original authors themselves annotated the errors in the generated textual units. They annotated cases of hallucinations, incorrect fact merging, omissions, redundancies and grammatical errors. In our reproduction study we attempt a reproduction on the same data samples, but narrow the scope to reproduce only the annotations of redundancy. We employ expert annotators to do this, as our common approach for reproduction prohibits reproduction authors to perform evaluations themselves.

This reproduction study was conducted as part of the ReproHum project[1] (Belz et al., 2023; Belz and Thomson, 2024), the aim of which is to build on existing work on recording properties of human evaluations datasheet-style (Shimorina and Belz, 2022) and assessing how close results from a reproduction study are to the original study (Belz et al., 2022), in order to systematically investigate what factors make human evaluations more—or less—reproducible. Taking part in this paper reproduction is a great opportunity to continue our own previous work in human evaluation (Jafaritazehjani et al., 2023, 2020; Klubička et al., 2018b,a; Klubička et al., 2017; Salton et al., 2014) and reproducibility (Klubička and Kelleher, 2023; Klubička and Fernández, 2018).

2. Original Study Design

In the original study the two authors themselves served as error annotators and annotated samples from two major triple-to-text datasets: WebNLG (Gardent et al., 2017; Castro Ferreira et al., 2020) and E2E (Novikova et al., 2017; Dušek et al., 2020). As their annotation interface they simply used a spreadsheet and noted the error counts in a column alongside the text samples. Each author was shown 300 text samples from each dataset and counted the number of errors in the sample. Notably, there was no overlap in samples, i.e. no text span was annotated by both authors, so inter-annotator agreement calculations were not possible. Given the authors served as annotators themselves and the task was deemed fairly straightforward, no annotation guidelines were developed or

[1] https://reprohum.github.io

Proceedings of the HumEval @LREC-COLING 2024, pages 163–198
Turin, Italy. May 21st, 2024. ©2024 European Language Resources Association (ELRA)

written, nor were the error categories explicitly defined, e.g. there was no common agreed upon understanding of what is redundancy. After the samples were annotated the authors discussed any edge cases and modified those annotations accordingly.

The authors made their model and data available in their GitHub repository[2]. However this does not include the final annotated data, which was instead shared via email with the ReproHum team upon request.

3. Reproduction Study Details

We used the exact same dataset used by Kasner and Dusek (2022), but in addition to focusing on a single quality criterion—redundancy—we also focused only on a single dataset, the E2E dataset (Novikova et al., 2017; Dušek et al., 2020). We copied the same 600 samples provided by the original authors, divided them between our two annotators and had each annotate 300 samples. Once the samples were annotated, we arranged for the annotators to meet and discuss edge cases. If they made any changes to their initial annotation, this was marked in a separate column next to the original annotation. Given this task was a simple integer count of occurrences in an output and involved no marking of text spans, there was no need to perform any postprocessing to obtain final annotations.

3.1. Evaluators

Our goal was to emulate the qualification of the original study's annotators, i.e. its authors who have experience in NLP research and are proficient in English. We thus internally recruited two colleagues: one a current PhD student of machine translation and one a recent PhD graduate in NLP.

Given there was no official annotation guide, we sent them brief instructions on how to perform the annotation in the spreadsheet, as well as their full dataset for annotation. They were told they can ask any practical questions should they arise, but should not communicate with each other or ask for opinions on how to annotate questionable instances until the later consolidation step, instead relying on their own judgement. The subsequent discussion of edge cases was also unmoderated: we simply organised a meeting and let the annotators discuss amongst themselves and come to a decision without any interference from our end.

In total, we estimated that the annotation would take around 5 hours of work, which turned out to be accurate. Given that the original authors also served as their own annotators, they were not directly paid for the annotation work. As in our case the annotators do not have the same incentives as the original authors—they will not get the satisfaction of a completed study and an authored publication as a result of the annotation—we instead compensated them financially. We followed the shared ReproHum procedure for calculating fair pay and paid them at a rate of €20/hour. This also exceeds the minimum wage in Ireland and would be considered fair pay for an annotation task.

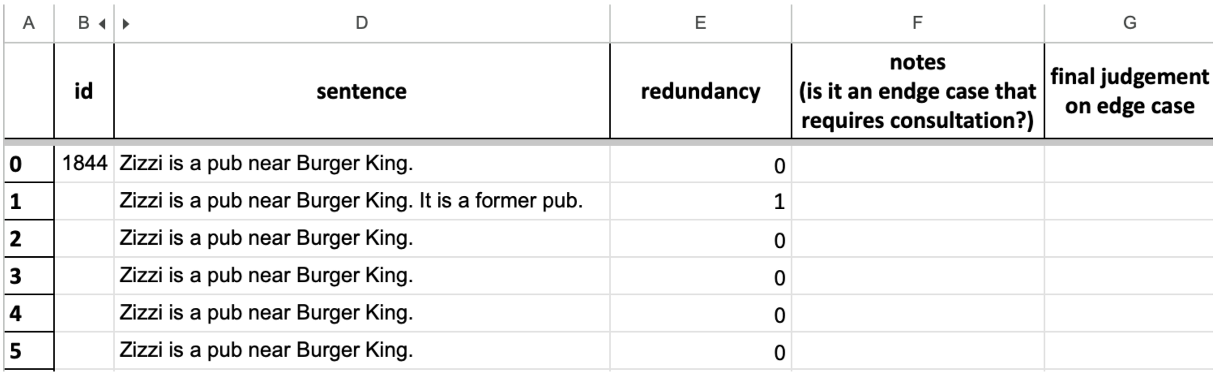
3.2. Differences

Any differences were fairly minor, and arguably the most impactful difference would be author involvement—the original study had the authors perform the error annotation, while in our case this did not align with our reproduction rules so we recruited external annotators.

Furthermore, based on the data provided by the original authors, they seem to have used an offline approach and worked in Microsoft Excel. In our case, we used the Google Sheets application and created a separate sheet that contained the data for each annotator individually. This approach made it straightforward to set up and more accessible to the annotators, as it was a familiar interface to them. The annotators were presented with the candidate text sample and three annotation columns (*redundancy count*, *edge case* and *final judgement*). Image 1 shows the annotation interface. This interface change is a minor difference, but arguably inconsequential.

Another seemingly small difference is the question of defining "redundancy". As there were no annotation guidelines in the original study, the authors presumably relied on their individual understanding, or perhaps reached a shared understanding while designing the study. This makes it difficult to make a decision on how to approach this question within our reproduction study—if we simply instruct the annotators to "count redundancies", and they return with a question of "what is redundancy", we must be able to say something. So after some internal discussion and communication with the ReproHum team, it was decided that prior to beginning the task, the annotators would be provided a definition of redundancy as follows: "a piece of information that has already been mentioned in the text". One could argue that this difference is also inconsequential, as people's intuition on what constitutes redundancy would be quite consistent, especially among academics who work in NLP. However minor differences are possible and providing a definition beforehand might smooth out that effect, so we still signpost this here in case it might have an impact.

[2] https://github.com/kasnerz/zeroshot-d2t-pipeline/

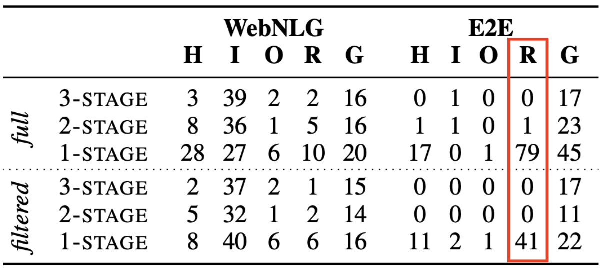


A	B	D	E	F	G
	id	sentence	redundancy	notes (is it an endge case that requires consultation?)	final judgement on edge case
0	1844	Zizzi is a pub near Burger King.	0		
1		Zizzi is a pub near Burger King. It is a former pub.	1		
2		Zizzi is a pub near Burger King.	0		
3		Zizzi is a pub near Burger King.	0		
4		Zizzi is a pub near Burger King.	0		
5		Zizzi is a pub near Burger King.	0		

Figure 1: Screenshot of the annotation interface shown to the evaluators.

4. Reproduction Results

The original paper developed 6 different data-to-text systems and when annotating redundancies they simply report the total error counts per system, as shown in Figure 2.



		WebNLG					E2E				
		H	I	O	R	G	H	I	O	R	G
full	3-STAGE	3	39	2	2	16	0	1	0	0	17
	2-STAGE	8	36	1	5	16	1	1	0	1	23
	1-STAGE	28	27	6	10	20	17	0	1	79	45
filtered	3-STAGE	2	37	2	1	15	0	0	0	0	17
	2-STAGE	5	32	1	2	14	0	0	0	0	11
	1-STAGE	8	40	6	6	16	11	2	1	41	22

Figure 2: Screenshot of the original paper's result table.

This numeric integer count is classified as a **Type I** result, as defined in the ReproHum reproduction guidelines. As such, we report side-by-side results from the original and repeat experiments in Table 1, both with initial counts and counts after the discussion step. It is interesting to note that the annotator discussion step yielded very few changes to their original assessments: while in total the annotators marked 22 samples as edge cases requiring discussion, they only changed the annotations of 3 samples after discussion. Due to this arguably inconsequential difference, we only calculate reproducibility assessments using the final error counts.

In order to quantify the reproducibility assessment for Type I results, we calculate the unbiased coefficient of variation for small samples (CV*) (Belz et al., 2022; Belz, 2022)[3], which we include in Table 1. Just by comparing the counts themselves it is already evident that there is a significant difference between our error counts and the originals, which is further supported by the high CV* values.

Labels	Original	Repro.	Final	CV*
1-stage	79	157	156	65.34
2-stage	1	11	11	166.17
3-stage	0	13	13	199.4
1-stage-F	41	85	84	68.59
2-stage-F	0	10	10	199.4
3-stage-F	0	10	9	199.4

Table 1: Redundancy error counts, comparing originally reported values, our own initially reproduced values, and the final values after the discussion step.

After some further analysis we note that the annotations can also be seen as **Type II** results, as they provide two distinct sets of numerical scores. It is thus possible to also quantify the reproducibility assessment via the Pearson or Spearman correlation coefficient. Given that our data is not ranked, but is simply a comparison of error counts, we calculate the Pearson correlation coefficient, which yields a result of **0.76**. This shows that the correlation between original and reproduced error counts is somewhere in the moderate-high range, indicating that the general trend is in fact being reproduced.

4.1. Findings Comparison

The original results presented in the paper by Kasner and Dusek (2022) relating to error annotation of redundancy find that the 1-stage model (which has to order the facts implicitly) tends to repeat the facts in the text, especially on the E2E dataset, which we also study. In their Appendix they also include examples showing how the 1-stage models add redundant information to the output.

We can clearly see in our results that this general trend has been reproduced: both 1-stage models have a dramatically higher number of redundancy occurrences when compared to 2-stage and 3-stage models. This is further supported by the high Pearson correlation coefficient. However it is surprising that our annotators were so much more

[3] Calculated using the provided Jupyter Notebook: https://github.com/asbelz/coeff-var

Scenario	Counts	Agreement
O=R=0	432	agree
O=0 R>0	76	disagree
O>0 R=0	3	disagree
O=R (>0)	37	agree
O<R (>0)	50	partial
O>R (>0)	2	partial

Table 2: Fine-grained counts of varying scenarios occurring when comparing the original and reproduced annotations, essentially showing the number of instances where annotators agree or disagree on the error counts.

liberal in annotating redundancy errors than the original authors, finding twice the amount of errors in 1-stage models. A brief analysis has shown that it was not a single annotator that contributed to the bulk of counted instances—both our annotators counted roughly (but not exactly) twice as many instances of redundancy as the original authors in their respective dataset splits.

This seemed unusual, so in order to gain more insight (and rule out any possible counting errors on our end) we analysed the annotation differences between the original (O) and reproduction (R) annotators. We identified six categories of interest: **a) O=R=0**, where O and R agree that there are 0 errors in the sample; **b) O=0 R>0**, where O counted 0 errors, while R counted >0; **c) O>0 R=0**, where O counted fewer errors than R (both >0); **d) O=R (>0)** where O and R counted the same number of errors, both >0; **e) O<R (>0)** where O counted fewer errors than R (both >0); and **f) O>R (>0)** where O counted more errors than R (both >0). We counted instances where these interactions occur and present these in Table 2.

In essence, the table provides a fine-grained view of the number of instances where the original and reproducing annotators agree or disagree in their error counts. We can see that in total they perfectly agree in 469 out of 600 instances. There is also "partial" agreement in 52 instances, where they agree there are some errors, but the error counts differ. They disagree a total of 79 times, i.e. one set of annotators found no errors, while the other set identified errors.

The disagreement scenario is particularly interesting, as it is the source of the large discrepancy in the error counts. The fine-grained look reveals that there are some instances where the original annotators found more errors than our annotators, however this number is quite low, totalling 5. It is significantly more frequent that our annotators have identified more errors than the original authors—a total of 126 instances—which makes up the majority of cases where our annotators identified a non-zero number of errors, far outweighing the 37 cases where both original and reproducing annotators agree on the exact number of errors.

This additional analysis likely rules out any simple counting or processing errors on our part, as there does not seem to be a clear function that consistently accounts for the discrepancies between the original and reproduced annotations: while very few, there are cases where the original annotators found more errors than the reproducing annotators, and there is a significant number of cases where they fully agree on the number of annotations. We wonder whether the inclusion of a "strict" definition of redundancy primed the annotators to overthink and be more critical of the content in the generated text. More likely, however, it indicates that the original and reproducing annotators had different annotation criteria.

We find support for this latter interpretation in a follow-up communication we had with our annotators. We reached out to them while analysing the results and writing up this report to ask if they would be willing to reflect and share any insights into their process, in hopes of explaining why they were prone to identifying a larger number of errors. In their feedback they noted that they approached the task by developing annotation heuristics that they aimed to apply consistently throughout the dataset. One annotator said that "*since the only guideline was to find repeated information, I set a standard I would follow and be consistent throughout the entire dataset*", with their biggest concern being sticking to their own established criteria. In regards to insight into their thought process whilst annotating, the same annotator said they were "*deconstructing sentences/segments into units and counting repetitions [of units]*". As examples, they provided the following: "*I remember clearly outlining 'fast food food' as a single repetition because 'fast food' was a unit and 'food' was another*". Another phrase of note was "*'low price range', where 'low' was a category and 'price range' was another. So if the phrase 'low price range' appeared twice, it would count as 2 repetitions as opposed to 3, as i did not subdivide it by word.*"

While we do not have such insights into the thought processes of the original annotators, this does elucidate the amount of subjective thought that goes into a task like this, and it is entirely possible that the original annotators had constructed a different framework for themselves when performing the annotation.

5. Conclusion

We successfully performed a reproduction study of redundancy error annotations on multiple data-to-text systems' outputs. We encountered no major challenges during the reproductions and sum-

Agree	Disagree
• general trend • 1-stage models exhibit much more redundancy than other models	• error counts • our annotators found more errors than original, as well as errors where original found none

Table 3: Summary table highlighting aspects of the study where our replication agreed and disagreed with the original experiment.

marise our key findings in Table 3. The findings point to the same general trends and conclusions as the original experiment. However, intriguingly, our annotators identified roughly twice as many redundancies in the dataset as the original authors—given how minor the differences were in our experiment implementation and execution, we found this puzzling, but cannot provide an answer as to why beyond speculation.

Acknowledgements

Many thanks to our annotators Somayeh Jafaritazehjani and João Lucas Cavalheiro Camargo for their time and effort. This research was financially supported by Science Foundation Ireland under Grant Agreements No. 13/RC/2106 and 13/RC/2106_P2 at the ADAPT SFI Research Centre at Technological University Dublin and Trinity College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme, and is co-funded under the European Regional Development Fund.

6. Bibliographical References

Anya Belz. 2022. A metrological perspective on reproducibility in NLP*. *Computational Linguistics*, 48(4):1125–1135.

Anya Belz, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2024. The 2024 repronlp shared task on reproducibility of evaluations in nlp: Overview and results. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*.

Anya Belz, Craig Thomson, and Ehud Reiter. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech Language*, 59:123–156.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Somayeh Jafaritazehjani, Gwénolé Lecorvé, Damien Lolive, and John D Kelleher. 2020. Style versus content: A distinction without a (learnable) difference? In *International Conference on Computational Linguistics*.

Somayeh Jafaritazehjani, Gwénolé Lecorvé, Damien Lolive, and John D Kelleher. 2023. Local or global: The variation in the encoding of style across sentiment and formality. In *International Conference on Artificial Neural Networks*, pages 492–504. Springer.

Zdeněk Kasner and Ondrej Dusek. 2022. Neural pipeline for zero-shot data-to-text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3914–3932, Dublin, Ireland. Association for Computational Linguistics.

Filip Klubička and John D. Kelleher. 2023. HumEval'23 reproduction report for paper 0040: Human evaluation of automatically detected over- and undertranslations. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 153–189, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Filip Klubička, Giancarlo D. Salton, and John D. Kelleher. 2018a. Is it worth it? budget-related evaluation metrics for model selection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Filip Klubička, Antonio Toral, and Víctor M Sánchez-Cartagena. 2018b. Quantitative fine-grained human evaluation of machine translation systems: a case study on english to croatian. *Machine Translation*, 32(3):195–215.

Filip Klubička, Antonio Toral Ruiz, and M. Víctor Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):121–132.

Filip Klubička and Raquel Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *Proceedings of 4REAL: 1st Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

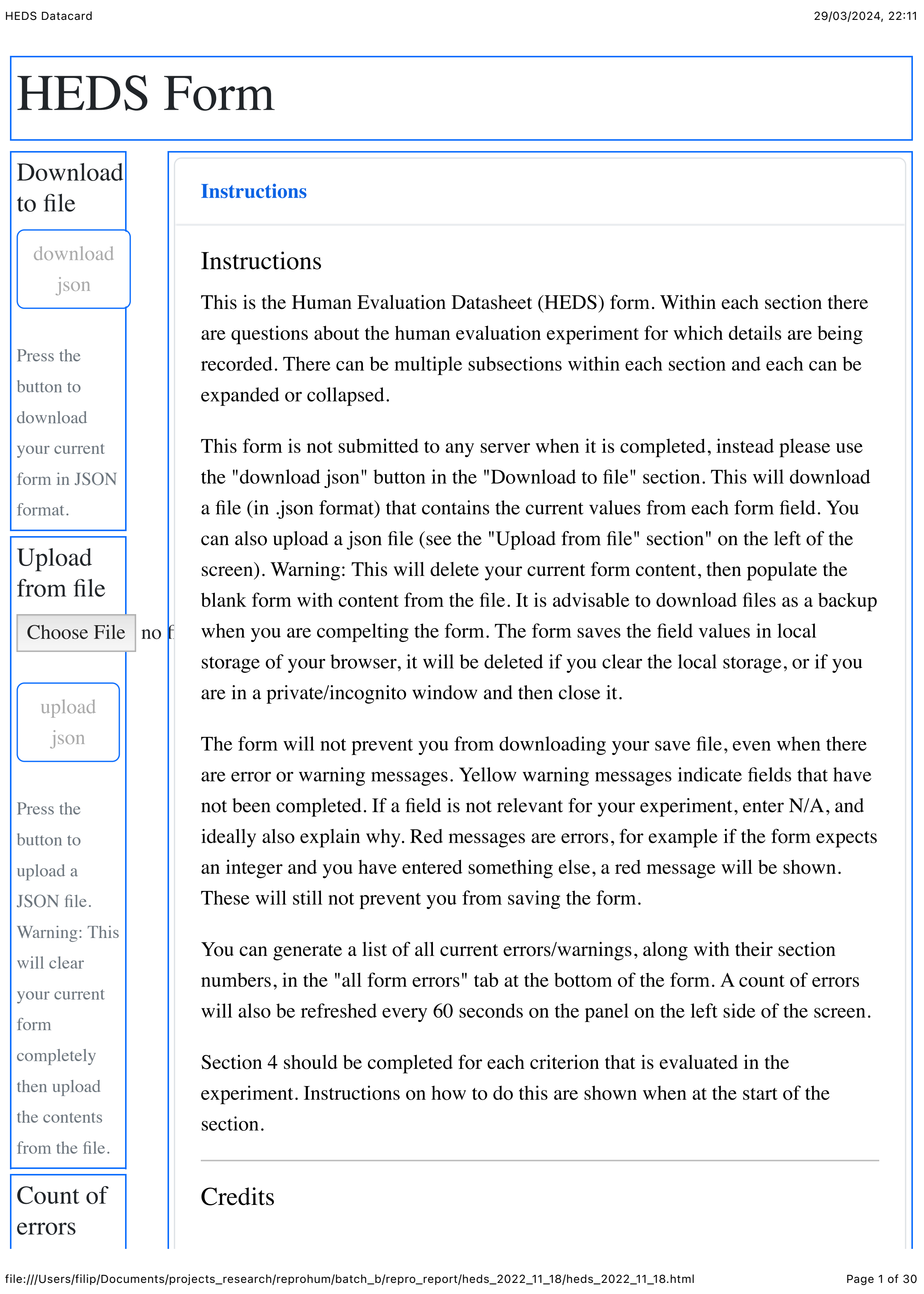
Giancarlo Salton, Robert Ross, and John Kelleher. 2014. Evaluation of a substitution method for idiom transformation in statistical machine translation. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 38–42, Gothenburg, Sweden. Association for Computational Linguistics.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

A. Appendix: Copy of the HEDS sheet

All project Human Evaluation DataSheets (HEDS) (Shimorina and Belz, 2022) can be found on the ReproHum GitHub page[4].

[4] `https://github.com/nlp-heds/repronlp2024`



HEDS Datacard 29/03/2024, 22:11

HEDS Form

Download to file

download json

Press the button to download your current form in JSON format.

Upload from file

Choose File no f

upload json

Press the button to upload a JSON file. Warning: This will clear your current form completely then upload the contents from the file.

Count of errors

Instructions

Instructions

This is the Human Evaluation Datasheet (HEDS) form. Within each section there are questions about the human evaluation experiment for which details are being recorded. There can be multiple subsections within each section and each can be expanded or collapsed.

This form is not submitted to any server when it is completed, instead please use the "download json" button in the "Download to file" section. This will download a file (in .json format) that contains the current values from each form field. You can also upload a json file (see the "Upload from file" section" on the left of the screen). Warning: This will delete your current form content, then populate the blank form with content from the file. It is advisable to download files as a backup when you are compelting the form. The form saves the field values in local storage of your browser, it will be deleted if you clear the local storage, or if you are in a private/incognito window and then close it.

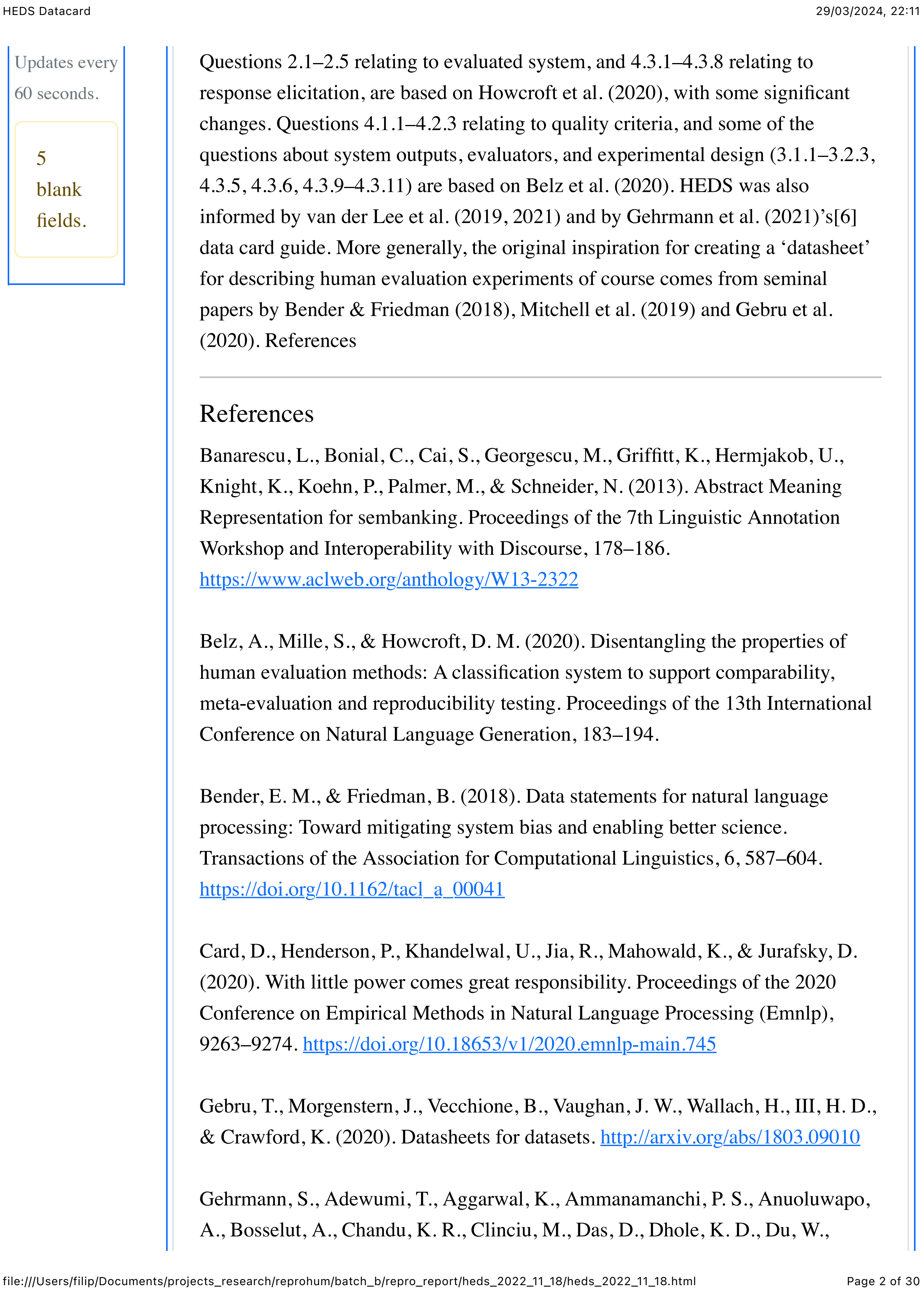
The form will not prevent you from downloading your save file, even when there are error or warning messages. Yellow warning messages indicate fields that have not been completed. If a field is not relevant for your experiment, enter N/A, and ideally also explain why. Red messages are errors, for example if the form expects an integer and you have entered something else, a red message will be shown. These will still not prevent you from saving the form.

You can generate a list of all current errors/warnings, along with their section numbers, in the "all form errors" tab at the bottom of the form. A count of errors will also be refreshed every 60 seconds on the panel on the left side of the screen.

Section 4 should be completed for each criterion that is evaluated in the experiment. Instructions on how to do this are shown when at the start of the section.

Credits

file:///Users/filip/Documents/projects_research/reprohum/batch_b/repro_report/heds_2022_11_18/heds_2022_11_18.html Page 1 of 30



Updates every 60 seconds.

5 blank fields.

Questions 2.1–2.5 relating to evaluated system, and 4.3.1–4.3.8 relating to response elicitation, are based on Howcroft et al. (2020), with some significant changes. Questions 4.1.1–4.2.3 relating to quality criteria, and some of the questions about system outputs, evaluators, and experimental design (3.1.1–3.2.3, 4.3.5, 4.3.6, 4.3.9–4.3.11) are based on Belz et al. (2020). HEDS was also informed by van der Lee et al. (2019, 2021) and by Gehrmann et al. (2021)'s[6] data card guide. More generally, the original inspiration for creating a 'datasheet' for describing human evaluation experiments of course comes from seminal papers by Bender & Friedman (2018), Mitchell et al. (2019) and Gebru et al. (2020). References

References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2013). Abstract Meaning Representation for sembanking. Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, 178–186. https://www.aclweb.org/anthology/W13-2322

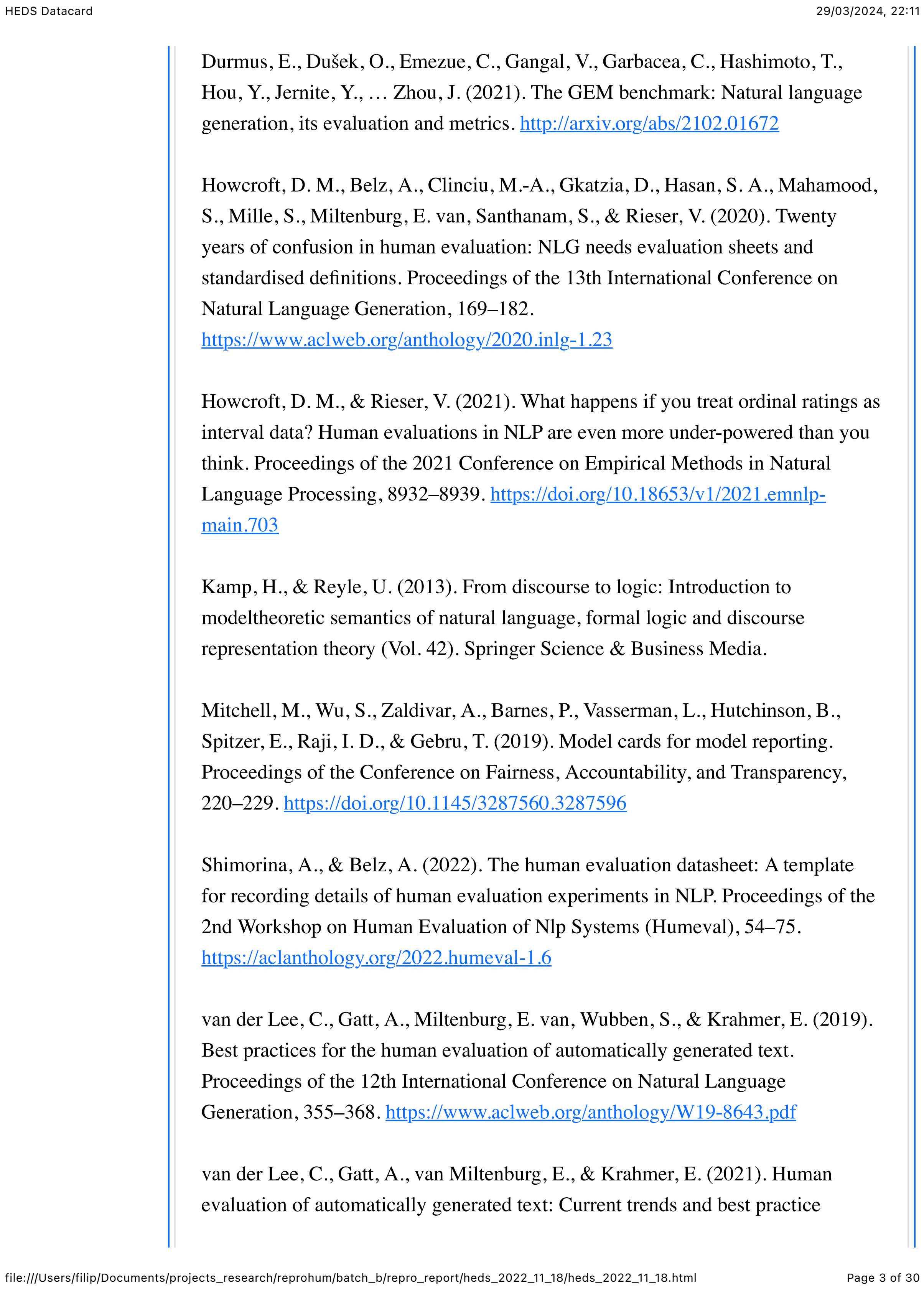
Belz, A., Mille, S., & Howcroft, D. M. (2020). Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. Proceedings of the 13th International Conference on Natural Language Generation, 183–194.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587–604. https://doi.org/10.1162/tacl_a_00041

Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., & Jurafsky, D. (2020). With little power comes great responsibility. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (Emnlp), 9263–9274. https://doi.org/10.18653/v1/2020.emnlp-main.745

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2020). Datasheets for datasets. http://arxiv.org/abs/1803.09010

Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Anuoluwapo, A., Bosselut, A., Chandu, K. R., Clinciu, M., Das, D., Dhole, K. D., Du, W.,



Durmus, E., Dušek, O., Emezue, C., Gangal, V., Garbacea, C., Hashimoto, T., Hou, Y., Jernite, Y., … Zhou, J. (2021). The GEM benchmark: Natural language generation, its evaluation and metrics. http://arxiv.org/abs/2102.01672

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., Miltenburg, E. van, Santhanam, S., & Rieser, V. (2020). Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. Proceedings of the 13th International Conference on Natural Language Generation, 169–182. https://www.aclweb.org/anthology/2020.inlg-1.23

Howcroft, D. M., & Rieser, V. (2021). What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 8932–8939. https://doi.org/10.18653/v1/2021.emnlp-main.703

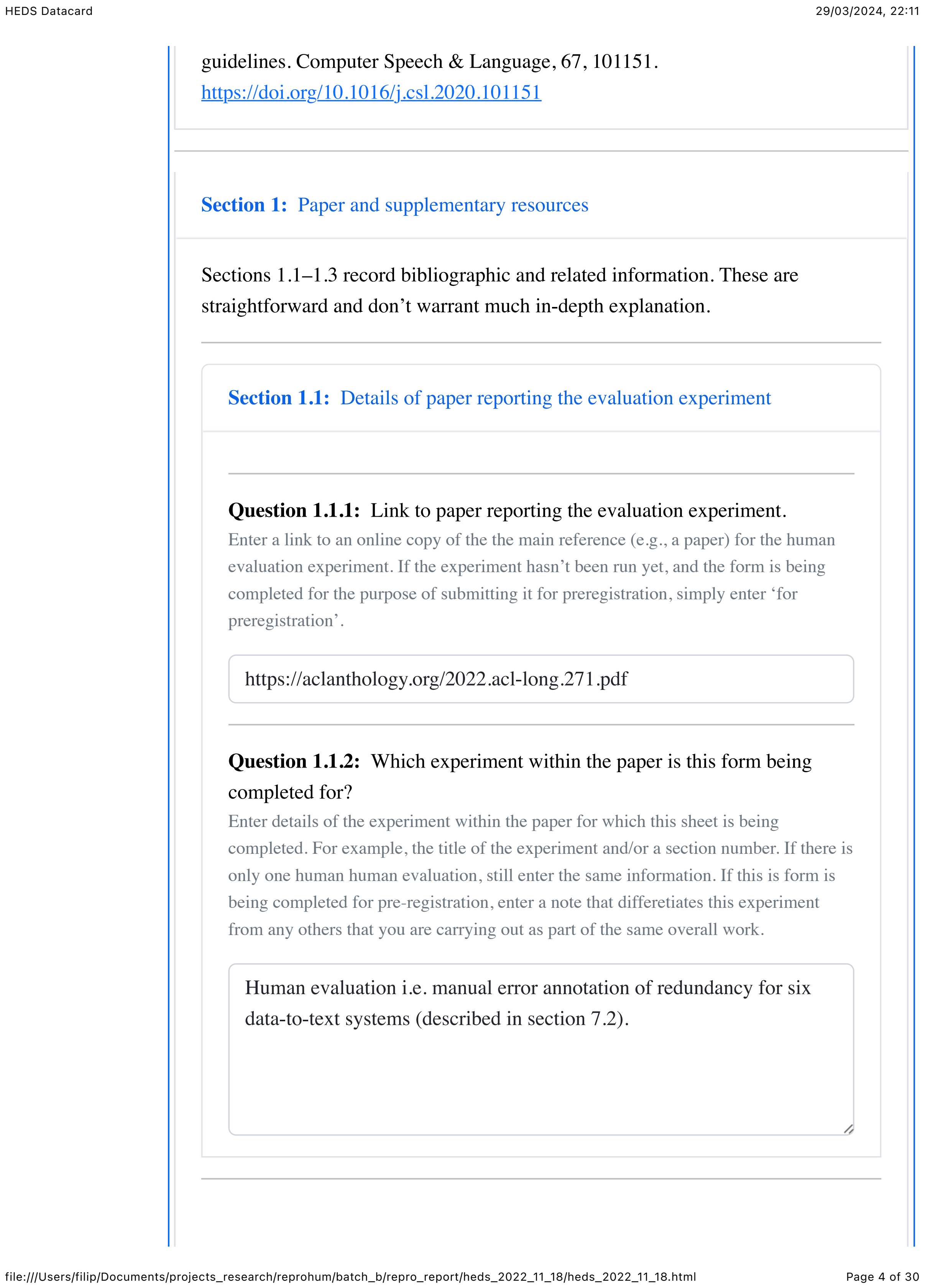
Kamp, H., & Reyle, U. (2013). From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory (Vol. 42). Springer Science & Business Media.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596

Shimorina, A., & Belz, A. (2022). The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. Proceedings of the 2nd Workshop on Human Evaluation of Nlp Systems (Humeval), 54–75. https://aclanthology.org/2022.humeval-1.6

van der Lee, C., Gatt, A., Miltenburg, E. van, Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. Proceedings of the 12th International Conference on Natural Language Generation, 355–368. https://www.aclweb.org/anthology/W19-8643.pdf

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice



guidelines. Computer Speech & Language, 67, 101151. https://doi.org/10.1016/j.csl.2020.101151

Section 1: Paper and supplementary resources

Sections 1.1–1.3 record bibliographic and related information. These are straightforward and don’t warrant much in-depth explanation.

Section 1.1: Details of paper reporting the evaluation experiment

Question 1.1.1: Link to paper reporting the evaluation experiment.

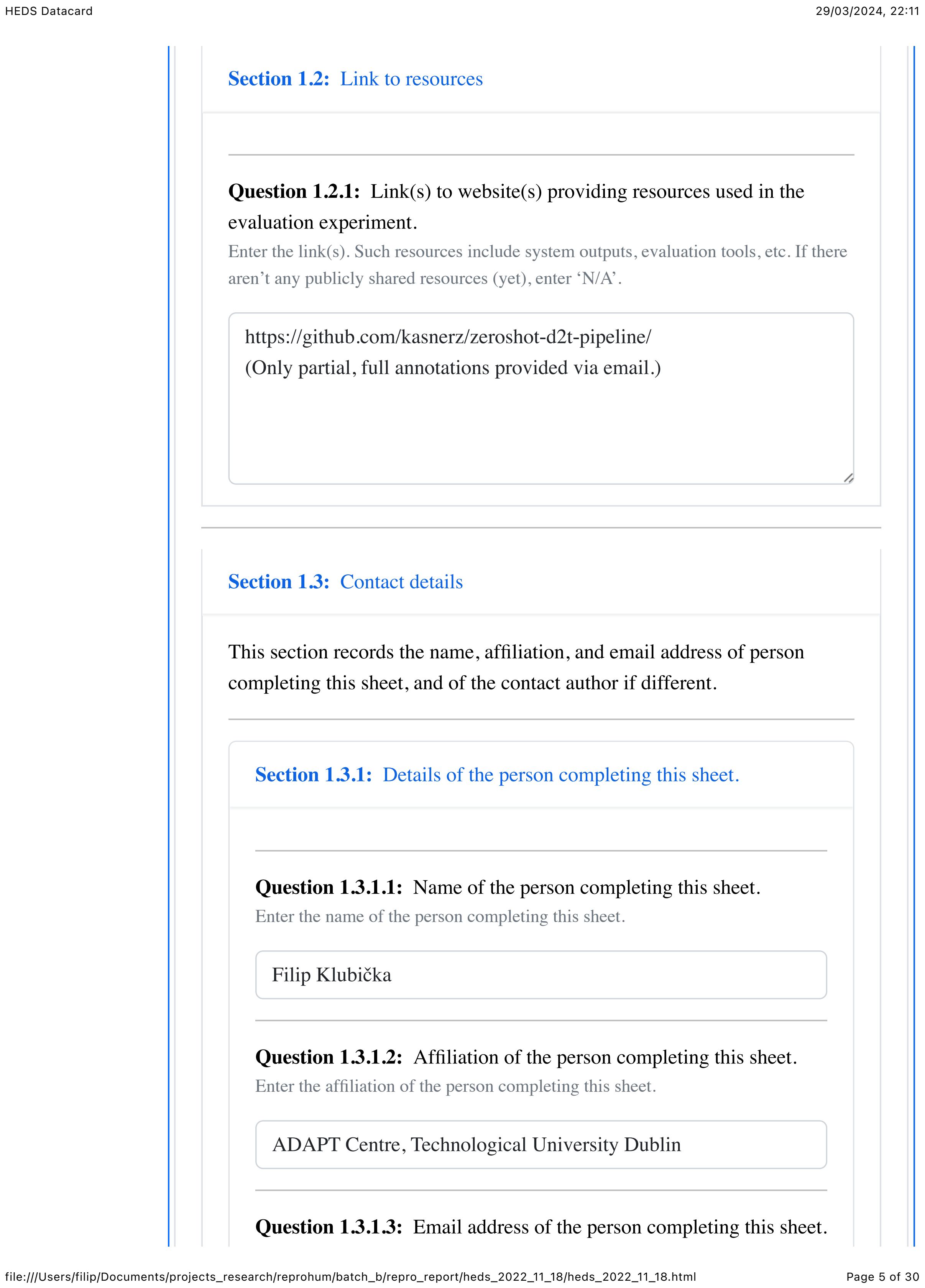
Enter a link to an online copy of the the main reference (e.g., a paper) for the human evaluation experiment. If the experiment hasn’t been run yet, and the form is being completed for the purpose of submitting it for preregistration, simply enter ‘for preregistration’.

https://aclanthology.org/2022.acl-long.271.pdf

Question 1.1.2: Which experiment within the paper is this form being completed for?

Enter details of the experiment within the paper for which this sheet is being completed. For example, the title of the experiment and/or a section number. If there is only one human human evaluation, still enter the same information. If this is form is being completed for pre-registration, enter a note that differetiates this experiment from any others that you are carrying out as part of the same overall work.

Human evaluation i.e. manual error annotation of redundancy for six data-to-text systems (described in section 7.2).



Section 1.2: Link to resources

Question 1.2.1: Link(s) to website(s) providing resources used in the evaluation experiment.

Enter the link(s). Such resources include system outputs, evaluation tools, etc. If there aren't any publicly shared resources (yet), enter 'N/A'.

https://github.com/kasnerz/zeroshot-d2t-pipeline/
(Only partial, full annotations provided via email.)

Section 1.3: Contact details

This section records the name, affiliation, and email address of person completing this sheet, and of the contact author if different.

Section 1.3.1: Details of the person completing this sheet.

Question 1.3.1.1: Name of the person completing this sheet.

Enter the name of the person completing this sheet.

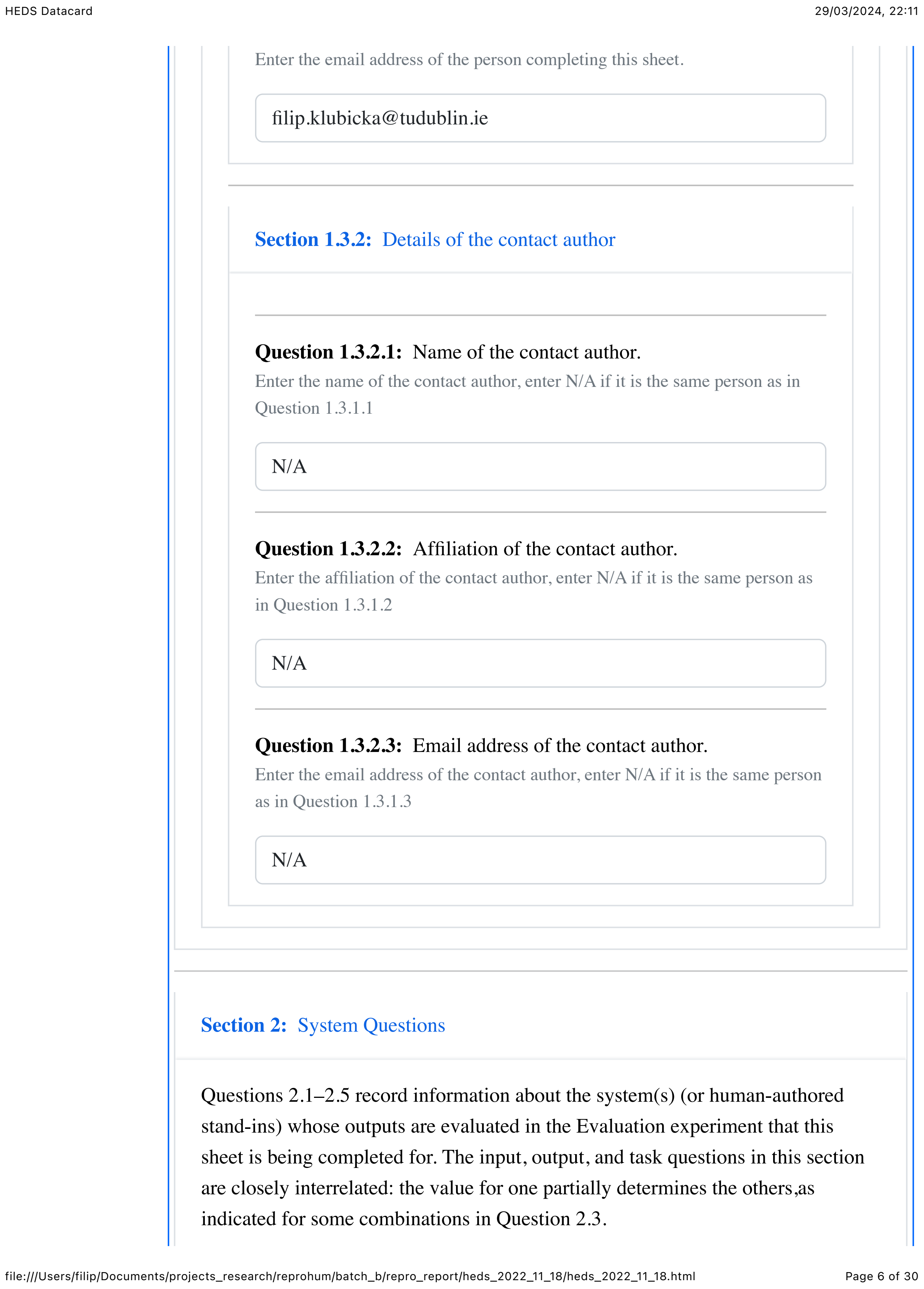
Filip Klubička

Question 1.3.1.2: Affiliation of the person completing this sheet.

Enter the affiliation of the person completing this sheet.

ADAPT Centre, Technological University Dublin

Question 1.3.1.3: Email address of the person completing this sheet.



Enter the email address of the person completing this sheet.

filip.klubicka@tudublin.ie

Section 1.3.2: Details of the contact author

Question 1.3.2.1: Name of the contact author.

Enter the name of the contact author, enter N/A if it is the same person as in Question 1.3.1.1

N/A

Question 1.3.2.2: Affiliation of the contact author.

Enter the affiliation of the contact author, enter N/A if it is the same person as in Question 1.3.1.2

N/A

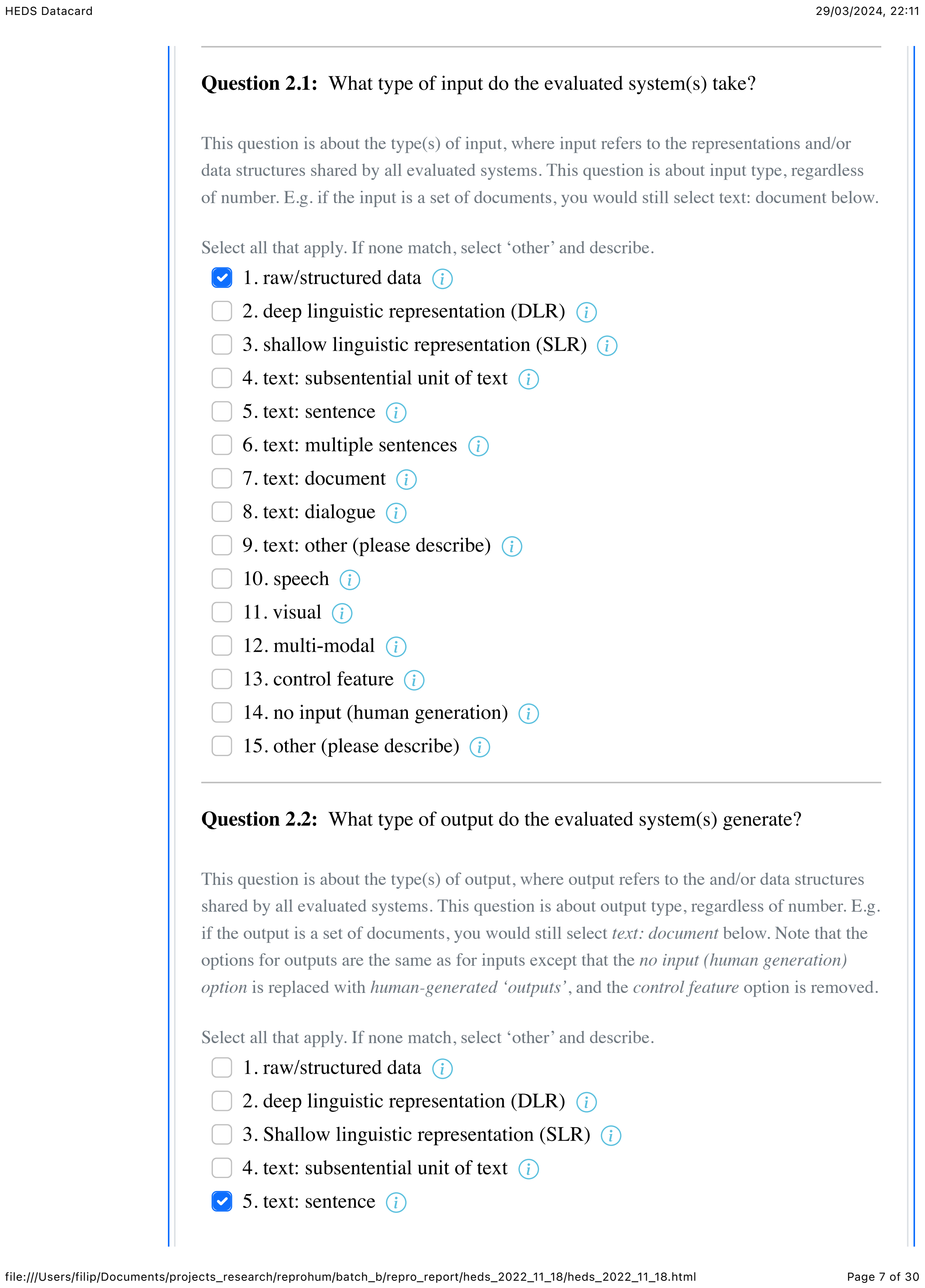
Question 1.3.2.3: Email address of the contact author.

Enter the email address of the contact author, enter N/A if it is the same person as in Question 1.3.1.3

N/A

Section 2: System Questions

Questions 2.1–2.5 record information about the system(s) (or human-authored stand-ins) whose outputs are evaluated in the Evaluation experiment that this sheet is being completed for. The input, output, and task questions in this section are closely interrelated: the value for one partially determines the others,as indicated for some combinations in Question 2.3.



Question 2.1: What type of input do the evaluated system(s) take?

This question is about the type(s) of input, where input refers to the representations and/or data structures shared by all evaluated systems. This question is about input type, regardless of number. E.g. if the input is a set of documents, you would still select text: document below.

Select all that apply. If none match, select ‘other’ and describe.

☑ 1. raw/structured data

□ 2. deep linguistic representation (DLR)

□ 3. shallow linguistic representation (SLR)

□ 4. text: subsentential unit of text

□ 5. text: sentence

□ 6. text: multiple sentences

□ 7. text: document

□ 8. text: dialogue

□ 9. text: other (please describe)

□ 10. speech

□ 11. visual

□ 12. multi-modal

□ 13. control feature

□ 14. no input (human generation)

□ 15. other (please describe)

Question 2.2: What type of output do the evaluated system(s) generate?

This question is about the type(s) of output, where output refers to the and/or data structures shared by all evaluated systems. This question is about output type, regardless of number. E.g. if the output is a set of documents, you would still select *text: document* below. Note that the options for outputs are the same as for inputs except that the *no input (human generation) option* is replaced with *human-generated ‘outputs’*, and the *control feature* option is removed.

Select all that apply. If none match, select ‘other’ and describe.

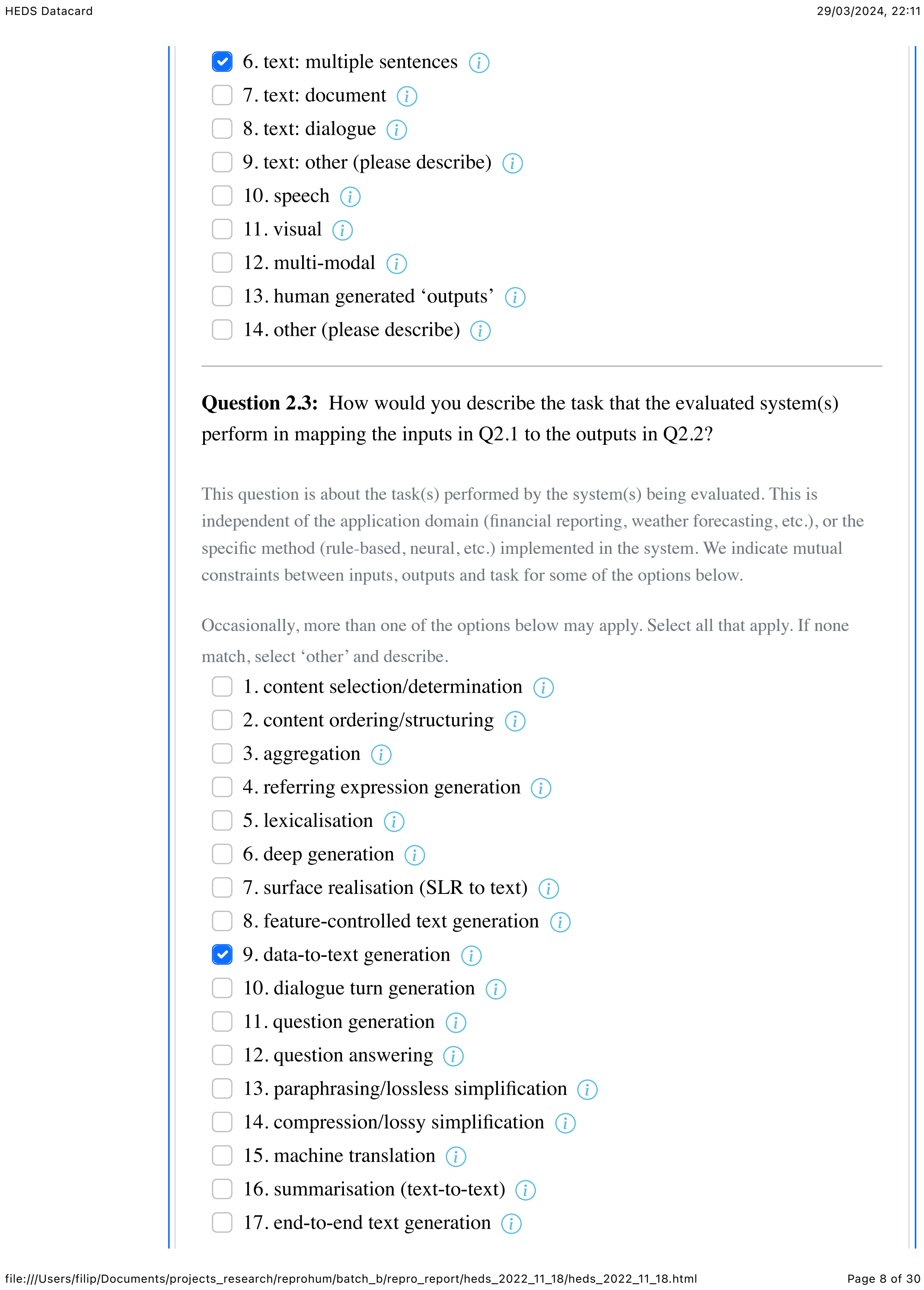
□ 1. raw/structured data

□ 2. deep linguistic representation (DLR)

□ 3. Shallow linguistic representation (SLR)

□ 4. text: subsentential unit of text

☑ 5. text: sentence



☑ 6. text: multiple sentences ⓘ
□ 7. text: document ⓘ
□ 8. text: dialogue ⓘ
□ 9. text: other (please describe) ⓘ
□ 10. speech ⓘ
□ 11. visual ⓘ
□ 12. multi-modal ⓘ
□ 13. human generated ‘outputs’ ⓘ
□ 14. other (please describe) ⓘ

Question 2.3: How would you describe the task that the evaluated system(s) perform in mapping the inputs in Q2.1 to the outputs in Q2.2?

This question is about the task(s) performed by the system(s) being evaluated. This is independent of the application domain (financial reporting, weather forecasting, etc.), or the specific method (rule-based, neural, etc.) implemented in the system. We indicate mutual constraints between inputs, outputs and task for some of the options below.

Occasionally, more than one of the options below may apply. Select all that apply. If none match, select ‘other’ and describe.

□ 1. content selection/determination ⓘ
□ 2. content ordering/structuring ⓘ
□ 3. aggregation ⓘ
□ 4. referring expression generation ⓘ
□ 5. lexicalisation ⓘ
□ 6. deep generation ⓘ
□ 7. surface realisation (SLR to text) ⓘ
□ 8. feature-controlled text generation ⓘ
☑ 9. data-to-text generation ⓘ
□ 10. dialogue turn generation ⓘ
□ 11. question generation ⓘ
□ 12. question answering ⓘ
□ 13. paraphrasing/lossless simplification ⓘ
□ 14. compression/lossy simplification ⓘ
□ 15. machine translation ⓘ
□ 16. summarisation (text-to-text) ⓘ
□ 17. end-to-end text generation ⓘ

□ 18. image/video description ⓘ
□ 19. post-editing/correction ⓘ
□ 20. other (please describe) ⓘ

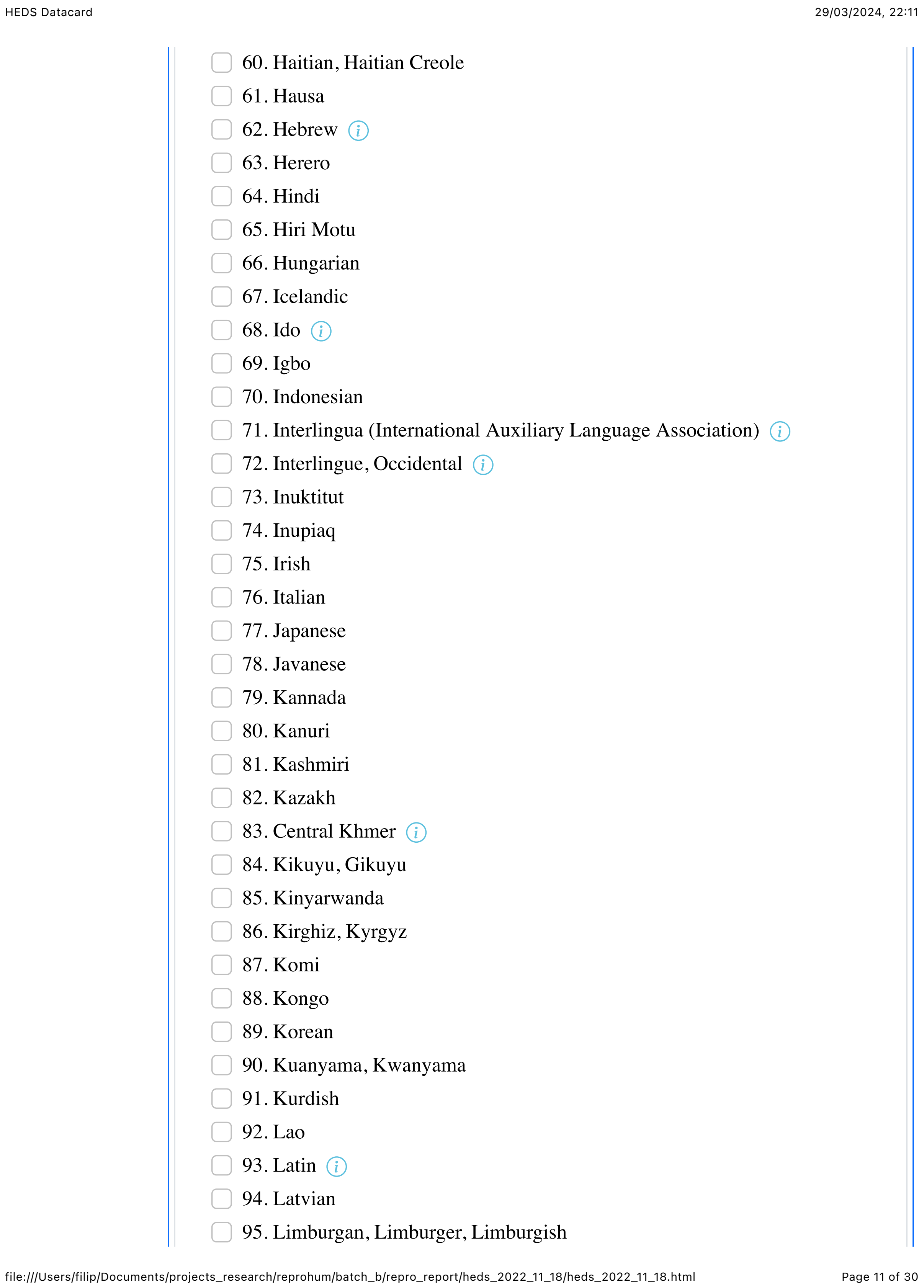
Question 2.4: What are the input languages that are used by the system?

This question is about the language(s) of the inputs accepted by the system(s) being evaluated. Select any language name(s) that apply, mapped to standardised full language names in ISO 639-1 (2019). E.g. English, Herero, Hindi. If no language is accepted as (part of) the input, select 'N/A'.

Select all that apply. If any languages you are using are not covered by this list, select 'other' and describe.

□ 1. Abkhazian ⓘ
□ 2. Afar
□ 3. Afrikaans
□ 4. Akan
□ 5. Albanian
□ 6. Amharic
□ 7. Arabic
□ 8. Aragonese
□ 9. Armenian
□ 10. Assamese
□ 11. Avaric ⓘ
□ 12. Avestan ⓘ
□ 13. Aymara
□ 14. Azerbaijani ⓘ
□ 15. Bambara
□ 16. Bashkir
□ 17. Basque
□ 18. Belarusian
□ 19. Bengali ⓘ
□ 20. Bislama ⓘ
□ 21. Bosnian
□ 22. Breton
□ 23. Bulgarian
□ 24. Burmese ⓘ

- □ 25. Catalan, Valencian
- □ 26. Chamorro
- □ 27. Chechen
- □ 28. Chichewa, Chewa, Nyanja
- □ 29. Chinese
- □ 30. Church Slavic, Old Slavonic, Church Slavonic, Old Bulgarian, Old Church Slavonic ⓘ
- □ 31. Chuvash
- □ 32. Cornish
- □ 33. Corsican
- □ 34. Cree
- □ 35. Croatian
- □ 36. Czech
- □ 37. Danish
- □ 38. Divehi, Dhivehi, Maldivian
- □ 39. Dutch, Flemish ⓘ
- □ 40. Dzongkha
- ☑ 41. English
- □ 42. Esperanto ⓘ
- □ 43. Estonian
- □ 44. Ewe
- □ 45. Faroese
- □ 46. Fijian
- □ 47. Finnish
- □ 48. French
- □ 49. Western Frisian ⓘ
- □ 50. Fulah ⓘ
- □ 51. Gaelic, Scottish Gaelic
- □ 52. Galician
- □ 53. Ganda
- □ 54. Georgian
- □ 55. German
- □ 56. Greek, Modern (1453–)
- □ 57. Kalaallisut, Greenlandic
- □ 58. Guarani
- □ 59. Gujarati



☐ 60. Haitian, Haitian Creole

☐ 61. Hausa

☐ 62. Hebrew

☐ 63. Herero

☐ 64. Hindi

☐ 65. Hiri Motu

☐ 66. Hungarian

☐ 67. Icelandic

☐ 68. Ido

☐ 69. Igbo

☐ 70. Indonesian

☐ 71. Interlingua (International Auxiliary Language Association)

☐ 72. Interlingue, Occidental

☐ 73. Inuktitut

☐ 74. Inupiaq

☐ 75. Irish

☐ 76. Italian

☐ 77. Japanese

☐ 78. Javanese

☐ 79. Kannada

☐ 80. Kanuri

☐ 81. Kashmiri

☐ 82. Kazakh

☐ 83. Central Khmer

☐ 84. Kikuyu, Gikuyu

☐ 85. Kinyarwanda

☐ 86. Kirghiz, Kyrgyz

☐ 87. Komi

☐ 88. Kongo

☐ 89. Korean

☐ 90. Kuanyama, Kwanyama

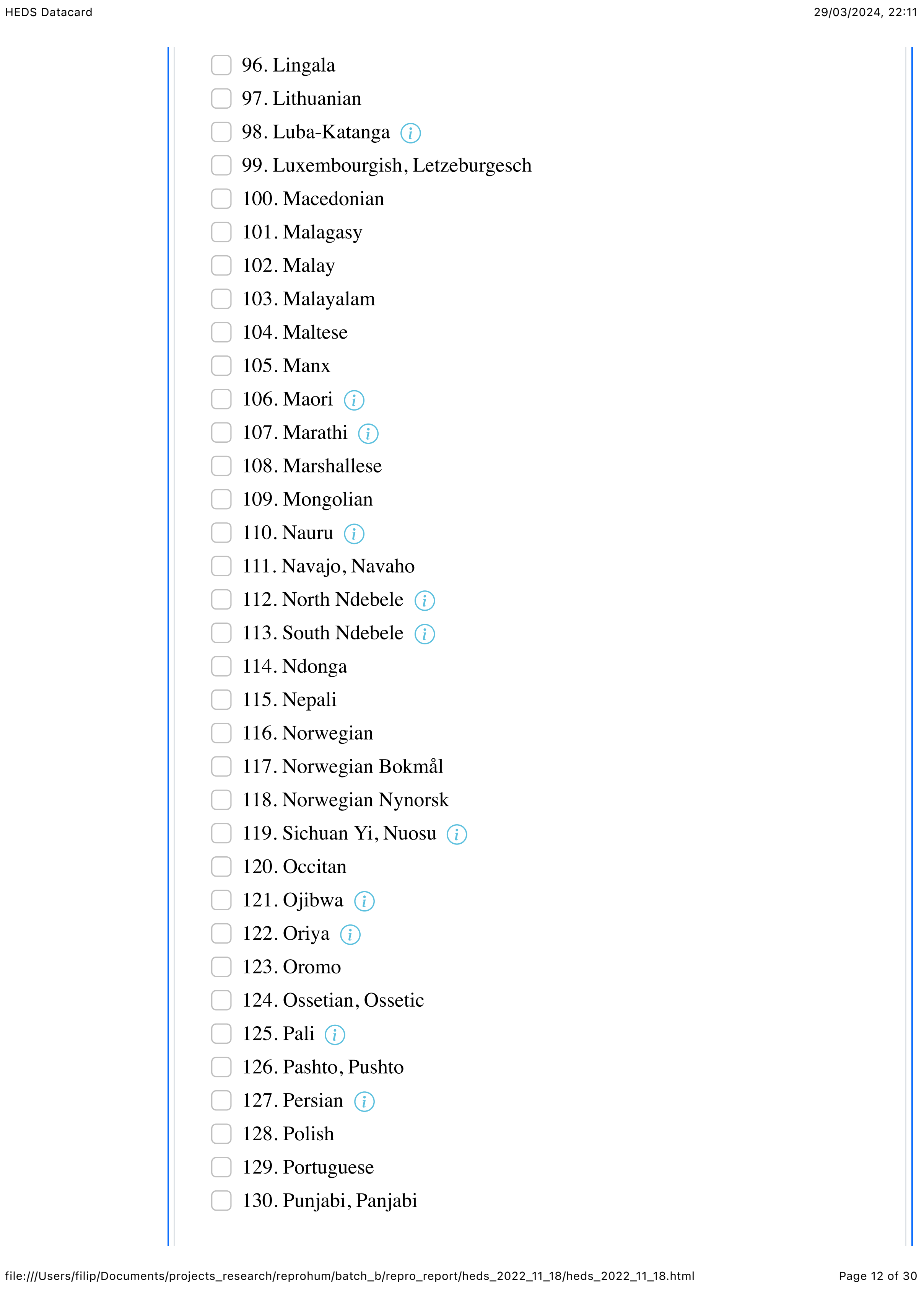
☐ 91. Kurdish

☐ 92. Lao

☐ 93. Latin

☐ 94. Latvian

☐ 95. Limburgan, Limburger, Limburgish



□ 96. Lingala

□ 97. Lithuanian

□ 98. Luba-Katanga ⓘ

□ 99. Luxembourgish, Letzeburgesch

□ 100. Macedonian

□ 101. Malagasy

□ 102. Malay

□ 103. Malayalam

□ 104. Maltese

□ 105. Manx

□ 106. Maori ⓘ

□ 107. Marathi ⓘ

□ 108. Marshallese

□ 109. Mongolian

□ 110. Nauru ⓘ

□ 111. Navajo, Navaho

□ 112. North Ndebele ⓘ

□ 113. South Ndebele ⓘ

□ 114. Ndonga

□ 115. Nepali

□ 116. Norwegian

□ 117. Norwegian Bokmål

□ 118. Norwegian Nynorsk

□ 119. Sichuan Yi, Nuosu ⓘ

□ 120. Occitan

□ 121. Ojibwa ⓘ

□ 122. Oriya ⓘ

□ 123. Oromo

□ 124. Ossetian, Ossetic

□ 125. Pali ⓘ

□ 126. Pashto, Pushto

□ 127. Persian ⓘ

□ 128. Polish

□ 129. Portuguese

□ 130. Punjabi, Panjabi

- □ 131. Quechua
- □ 132. Romanian, Moldavian, Moldovan
- □ 133. Romansh
- □ 134. Rundi ⓘ
- □ 135. Russian
- □ 136. Northern Sami
- □ 137. Samoan
- □ 138. Sango
- □ 139. Sanskrit ⓘ
- □ 140. Sardinian
- □ 141. Serbian
- □ 142. Shona
- □ 143. Sindhi
- □ 144. Sinhala, Sinhalese
- □ 145. Slovak
- □ 146. Slovenian ⓘ
- □ 147. Somali
- □ 148. Southern Sotho
- □ 149. Spanish, Castilian
- □ 150. Sundanese
- □ 151. Swahili
- □ 152. Swati ⓘ
- □ 153. Swedish
- □ 154. Tagalog
- □ 155. Tahitian ⓘ
- □ 156. Tajik
- □ 157. Tamil
- □ 158. Tatar
- □ 159. Telugu
- □ 160. Thai
- □ 161. Tibetan ⓘ
- □ 162. Tigrinya
- □ 163. Tonga (Tonga Islands) ⓘ
- □ 164. Tsonga
- □ 165. Tswana
- □ 166. Turkish

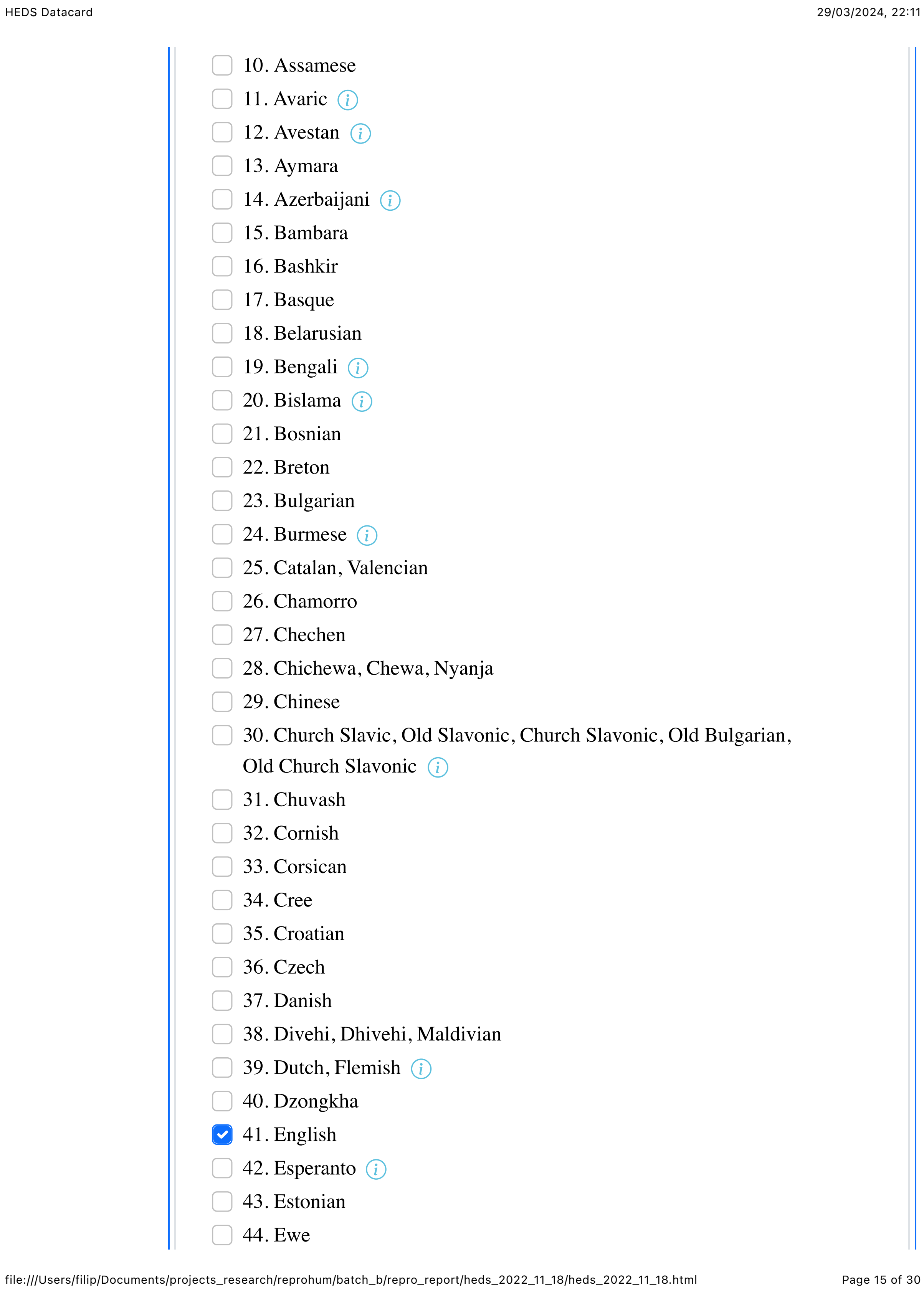
□ 167. Turkmen
□ 168. Twi
□ 169. Uighur, Uyghur
□ 170. Ukrainian
□ 171. Urdu
□ 172. Uzbek
□ 173. Venda
□ 174. Vietnamese
□ 175. Volapük ⓘ
□ 176. Walloon
□ 177. Welsh
□ 178. Wolof
□ 179. Xhosa
□ 180. Yiddish
□ 181. Yoruba
□ 182. Zhuang, Chuang
□ 183. Zulu
□ 184. Other (please describe) ⓘ
□ 185. N/A (please describe) ⓘ

Question 2.5: What are the output languages that are used by the system?

This field question the language(s) of the outputs generated by the system(s) being evaluated. Select any language name(s) that apply, mapped to standardised full language names in ISO 639-1 (2019). E.g. English, Herero, Hindi. If no language is generated, select 'N/A'.

Select all that apply. If any languages you are using are not covered by this list, select 'other' and describe.

□ 1. Abkhazian ⓘ
□ 2. Afar
□ 3. Afrikaans
□ 4. Akan
□ 5. Albanian
□ 6. Amharic
□ 7. Arabic
□ 8. Aragonese
□ 9. Armenian



□ 10. Assamese

□ 11. Avaric ⓘ

□ 12. Avestan ⓘ

□ 13. Aymara

□ 14. Azerbaijani ⓘ

□ 15. Bambara

□ 16. Bashkir

□ 17. Basque

□ 18. Belarusian

□ 19. Bengali ⓘ

□ 20. Bislama ⓘ

□ 21. Bosnian

□ 22. Breton

□ 23. Bulgarian

□ 24. Burmese ⓘ

□ 25. Catalan, Valencian

□ 26. Chamorro

□ 27. Chechen

□ 28. Chichewa, Chewa, Nyanja

□ 29. Chinese

□ 30. Church Slavic, Old Slavonic, Church Slavonic, Old Bulgarian, Old Church Slavonic ⓘ

□ 31. Chuvash

□ 32. Cornish

□ 33. Corsican

□ 34. Cree

□ 35. Croatian

□ 36. Czech

□ 37. Danish

□ 38. Divehi, Dhivehi, Maldivian

□ 39. Dutch, Flemish ⓘ

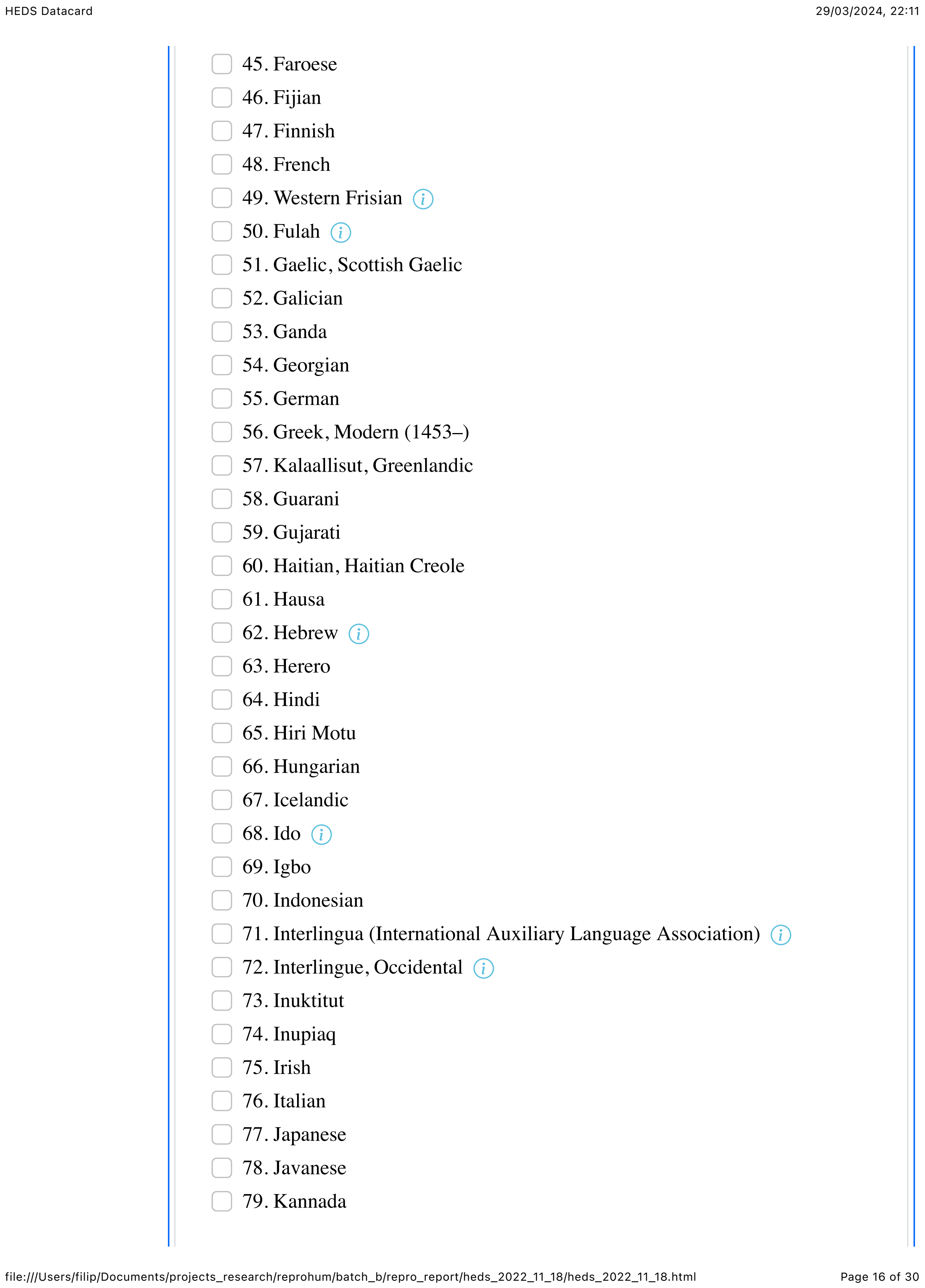
□ 40. Dzongkha

☑ 41. English

□ 42. Esperanto ⓘ

□ 43. Estonian

□ 44. Ewe



HEDS Datacard 29/03/2024, 22:11

- □ 45. Faroese
- □ 46. Fijian
- □ 47. Finnish
- □ 48. French
- □ 49. Western Frisian ⓘ
- □ 50. Fulah ⓘ
- □ 51. Gaelic, Scottish Gaelic
- □ 52. Galician
- □ 53. Ganda
- □ 54. Georgian
- □ 55. German
- □ 56. Greek, Modern (1453–)
- □ 57. Kalaallisut, Greenlandic
- □ 58. Guarani
- □ 59. Gujarati
- □ 60. Haitian, Haitian Creole
- □ 61. Hausa
- □ 62. Hebrew ⓘ
- □ 63. Herero
- □ 64. Hindi
- □ 65. Hiri Motu
- □ 66. Hungarian
- □ 67. Icelandic
- □ 68. Ido ⓘ
- □ 69. Igbo
- □ 70. Indonesian
- □ 71. Interlingua (International Auxiliary Language Association) ⓘ
- □ 72. Interlingue, Occidental ⓘ
- □ 73. Inuktitut
- □ 74. Inupiaq
- □ 75. Irish
- □ 76. Italian
- □ 77. Japanese
- □ 78. Javanese
- □ 79. Kannada

file:///Users/filip/Documents/projects_research/reprohum/batch_b/repro_report/heds_2022_11_18/heds_2022_11_18.html Page 16 of 30

- □ 80. Kanuri
- □ 81. Kashmiri
- □ 82. Kazakh
- □ 83. Central Khmer ⓘ
- □ 84. Kikuyu, Gikuyu
- □ 85. Kinyarwanda
- □ 86. Kirghiz, Kyrgyz
- □ 87. Komi
- □ 88. Kongo
- □ 89. Korean
- □ 90. Kuanyama, Kwanyama
- □ 91. Kurdish
- □ 92. Lao
- □ 93. Latin ⓘ
- □ 94. Latvian
- □ 95. Limburgan, Limburger, Limburgish
- □ 96. Lingala
- □ 97. Lithuanian
- □ 98. Luba-Katanga ⓘ
- □ 99. Luxembourgish, Letzeburgesch
- □ 100. Macedonian
- □ 101. Malagasy
- □ 102. Malay
- □ 103. Malayalam
- □ 104. Maltese
- □ 105. Manx
- □ 106. Maori ⓘ
- □ 107. Marathi ⓘ
- □ 108. Marshallese
- □ 109. Mongolian
- □ 110. Nauru ⓘ
- □ 111. Navajo, Navaho
- □ 112. North Ndebele ⓘ
- □ 113. South Ndebele ⓘ
- □ 114. Ndonga
- □ 115. Nepali

□ 116. Norwegian

□ 117. Norwegian Bokmål

□ 118. Norwegian Nynorsk

□ 119. Sichuan Yi, Nuosu ⓘ

□ 120. Occitan

□ 121. Ojibwa ⓘ

□ 122. Oriya ⓘ

□ 123. Oromo

□ 124. Ossetian, Ossetic

□ 125. Pali ⓘ

□ 126. Pashto, Pushto

□ 127. Persian ⓘ

□ 128. Polish

□ 129. Portuguese

□ 130. Punjabi, Panjabi

□ 131. Quechua

□ 132. Romanian, Moldavian, Moldovan

□ 133. Romansh

□ 134. Rundi ⓘ

□ 135. Russian

□ 136. Northern Sami

□ 137. Samoan

□ 138. Sango

□ 139. Sanskrit ⓘ

□ 140. Sardinian

□ 141. Serbian

□ 142. Shona

□ 143. Sindhi

□ 144. Sinhala, Sinhalese

□ 145. Slovak

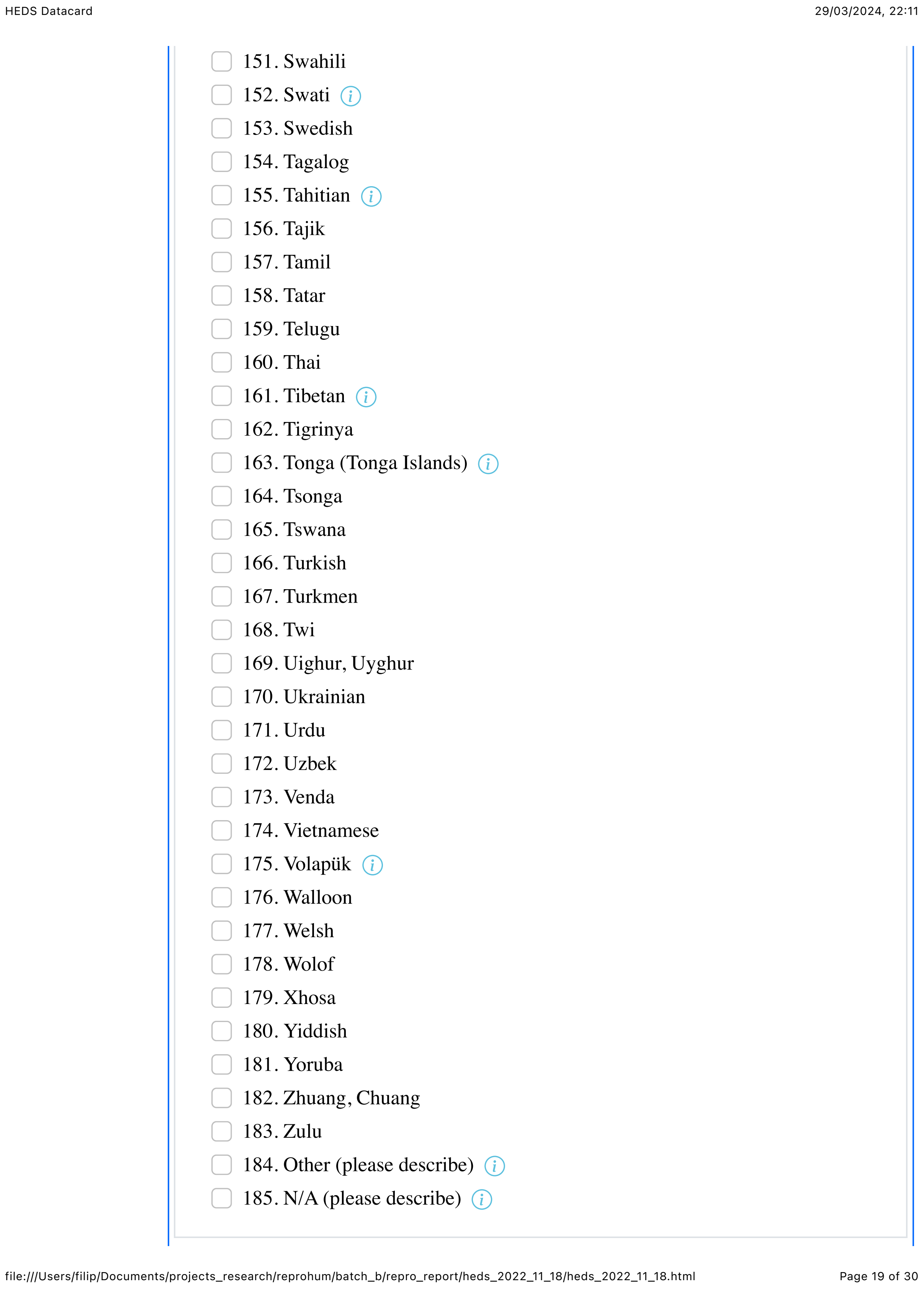
□ 146. Slovenian ⓘ

□ 147. Somali

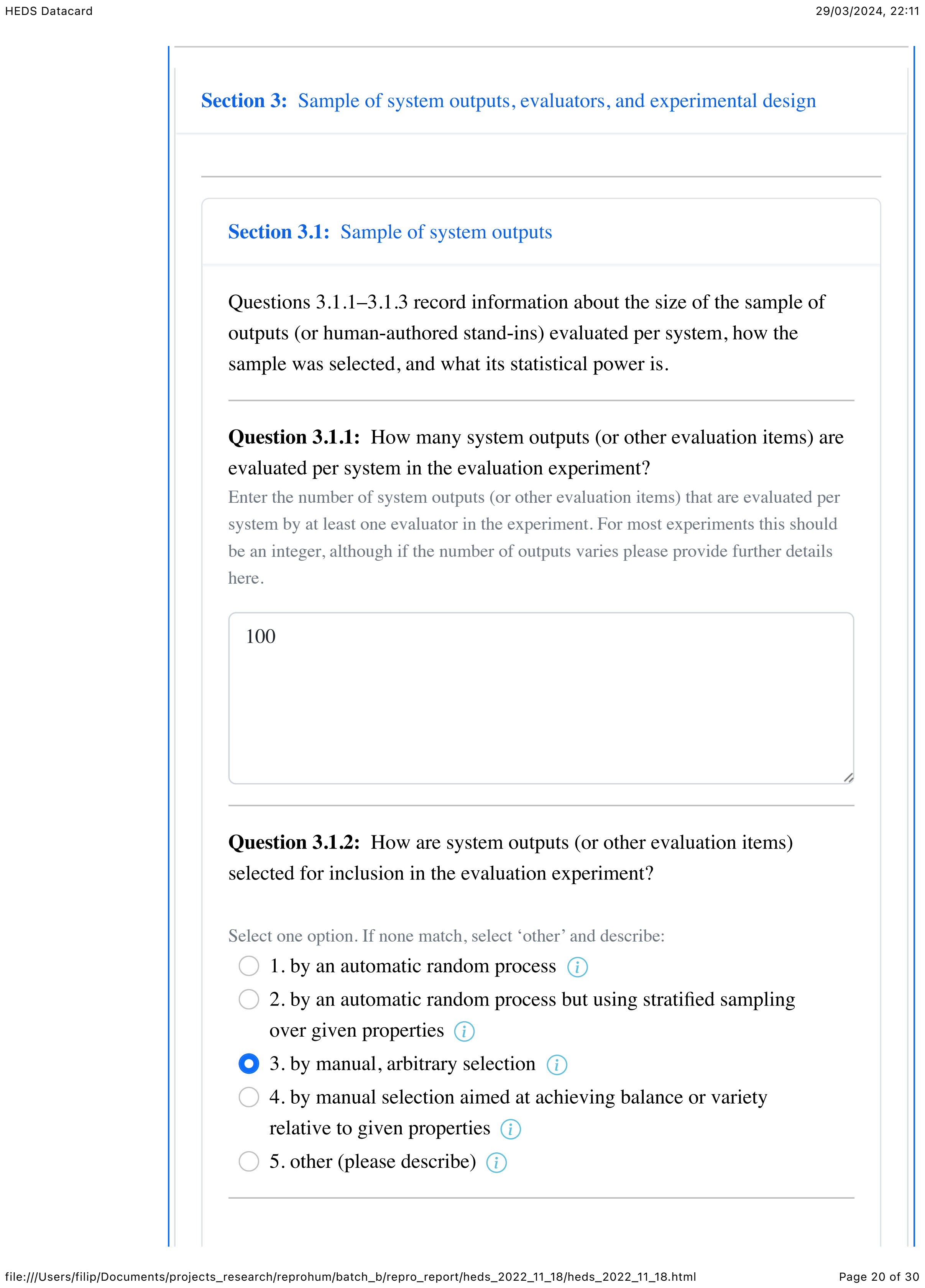
□ 148. Southern Sotho

□ 149. Spanish, Castilian

□ 150. Sundanese



- □ 151. Swahili
- □ 152. Swati ⓘ
- □ 153. Swedish
- □ 154. Tagalog
- □ 155. Tahitian ⓘ
- □ 156. Tajik
- □ 157. Tamil
- □ 158. Tatar
- □ 159. Telugu
- □ 160. Thai
- □ 161. Tibetan ⓘ
- □ 162. Tigrinya
- □ 163. Tonga (Tonga Islands) ⓘ
- □ 164. Tsonga
- □ 165. Tswana
- □ 166. Turkish
- □ 167. Turkmen
- □ 168. Twi
- □ 169. Uighur, Uyghur
- □ 170. Ukrainian
- □ 171. Urdu
- □ 172. Uzbek
- □ 173. Venda
- □ 174. Vietnamese
- □ 175. Volapük ⓘ
- □ 176. Walloon
- □ 177. Welsh
- □ 178. Wolof
- □ 179. Xhosa
- □ 180. Yiddish
- □ 181. Yoruba
- □ 182. Zhuang, Chuang
- □ 183. Zulu
- □ 184. Other (please describe) ⓘ
- □ 185. N/A (please describe) ⓘ



Section 3: Sample of system outputs, evaluators, and experimental design

Section 3.1: Sample of system outputs

Questions 3.1.1–3.1.3 record information about the size of the sample of outputs (or human-authored stand-ins) evaluated per system, how the sample was selected, and what its statistical power is.

Question 3.1.1: How many system outputs (or other evaluation items) are evaluated per system in the evaluation experiment?

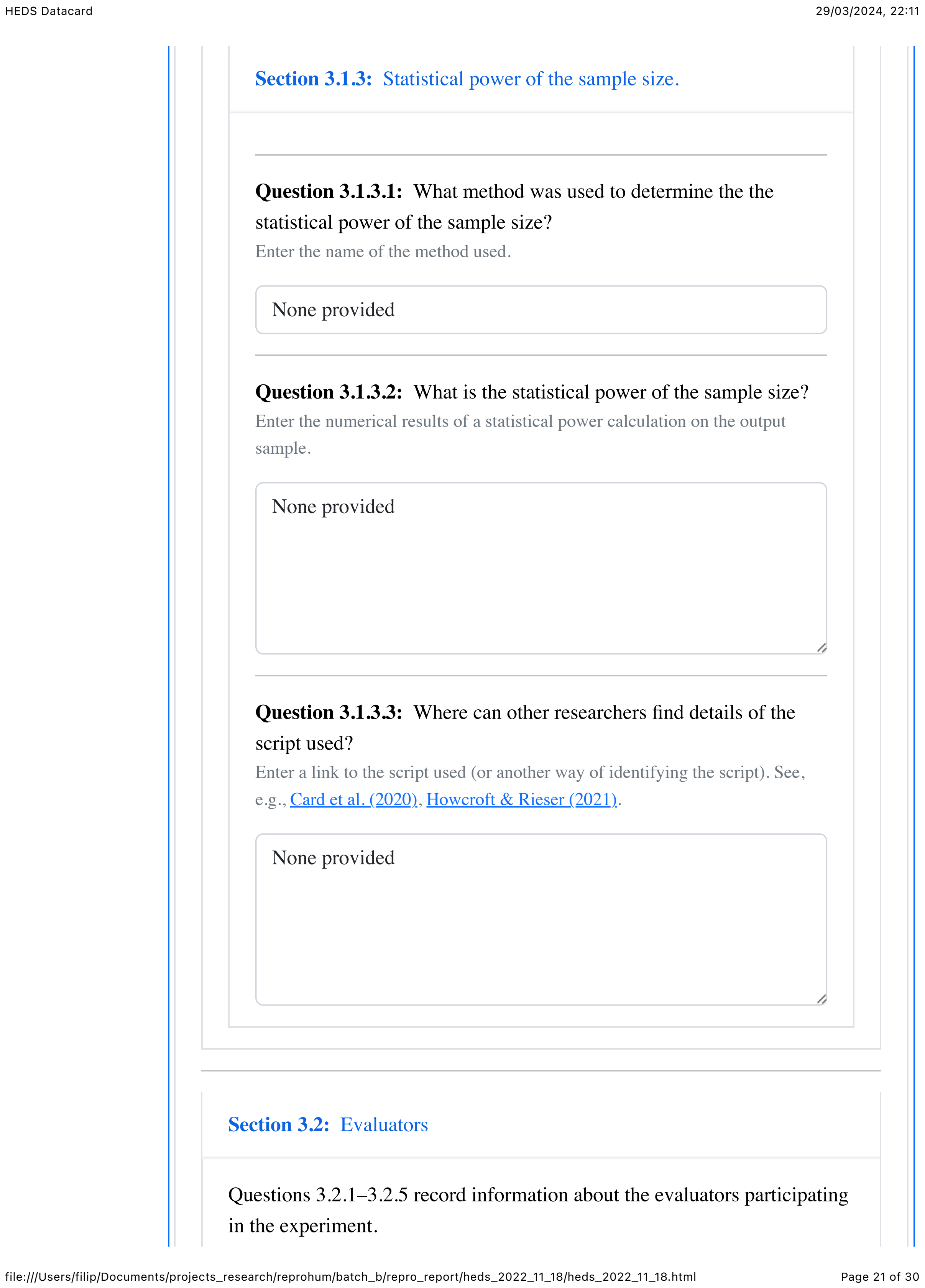
Enter the number of system outputs (or other evaluation items) that are evaluated per system by at least one evaluator in the experiment. For most experiments this should be an integer, although if the number of outputs varies please provide further details here.

100

Question 3.1.2: How are system outputs (or other evaluation items) selected for inclusion in the evaluation experiment?

Select one option. If none match, select 'other' and describe:

- ○ 1. by an automatic random process ⓘ
- ○ 2. by an automatic random process but using stratified sampling over given properties ⓘ
- ● 3. by manual, arbitrary selection ⓘ
- ○ 4. by manual selection aimed at achieving balance or variety relative to given properties ⓘ
- ○ 5. other (please describe) ⓘ



Section 3.1.3: Statistical power of the sample size.

Question 3.1.3.1: What method was used to determine the the statistical power of the sample size?

Enter the name of the method used.

None provided

Question 3.1.3.2: What is the statistical power of the sample size?

Enter the numerical results of a statistical power calculation on the output sample.

None provided

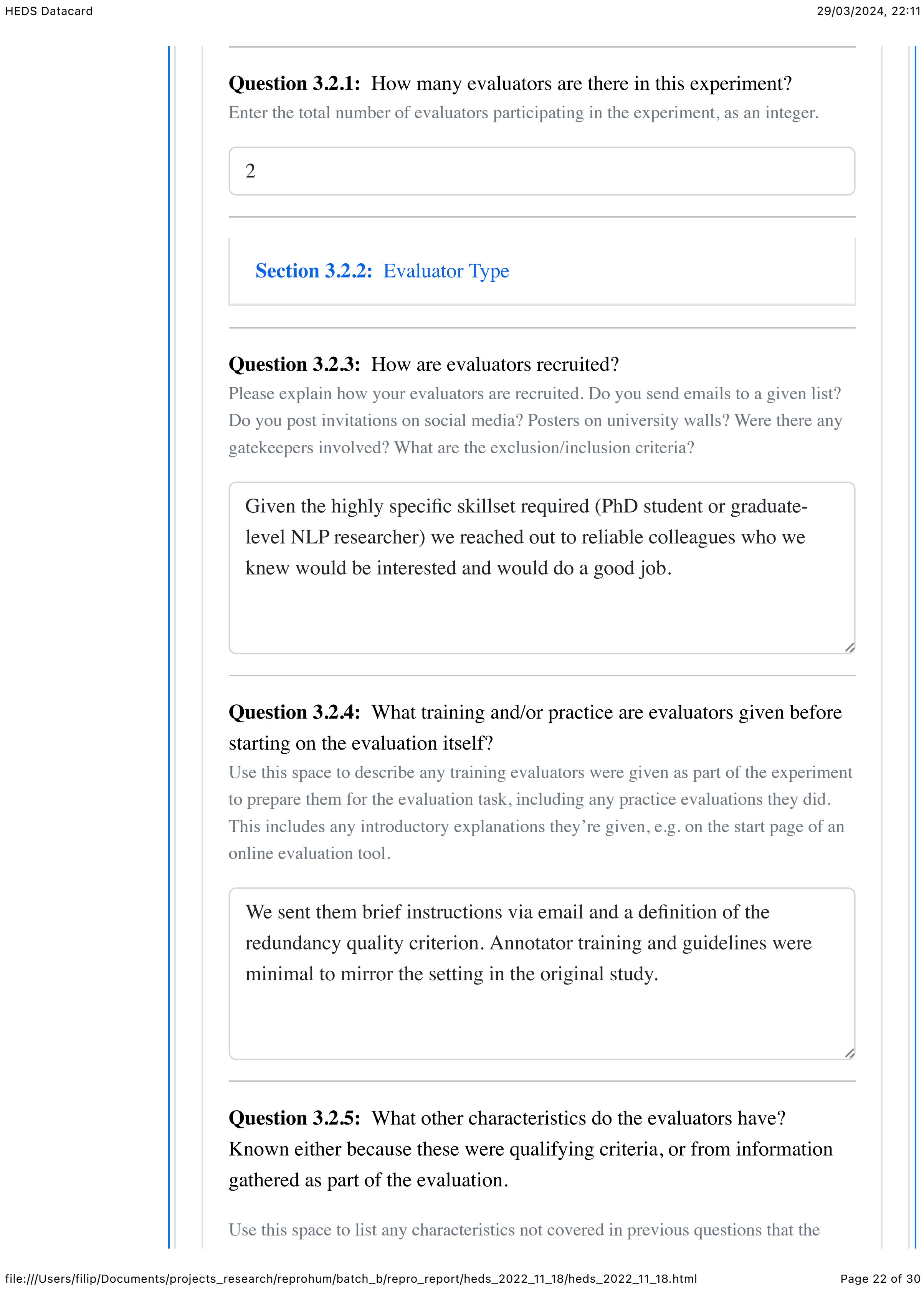
Question 3.1.3.3: Where can other researchers find details of the script used?

Enter a link to the script used (or another way of identifying the script). See, e.g., Card et al. (2020), Howcroft & Rieser (2021).

None provided

Section 3.2: Evaluators

Questions 3.2.1–3.2.5 record information about the evaluators participating in the experiment.



Question 3.2.1: How many evaluators are there in this experiment?

Enter the total number of evaluators participating in the experiment, as an integer.

2

Section 3.2.2: Evaluator Type

Question 3.2.3: How are evaluators recruited?

Please explain how your evaluators are recruited. Do you send emails to a given list? Do you post invitations on social media? Posters on university walls? Were there any gatekeepers involved? What are the exclusion/inclusion criteria?

Given the highly specific skillset required (PhD student or graduate-level NLP researcher) we reached out to reliable colleagues who we knew would be interested and would do a good job.

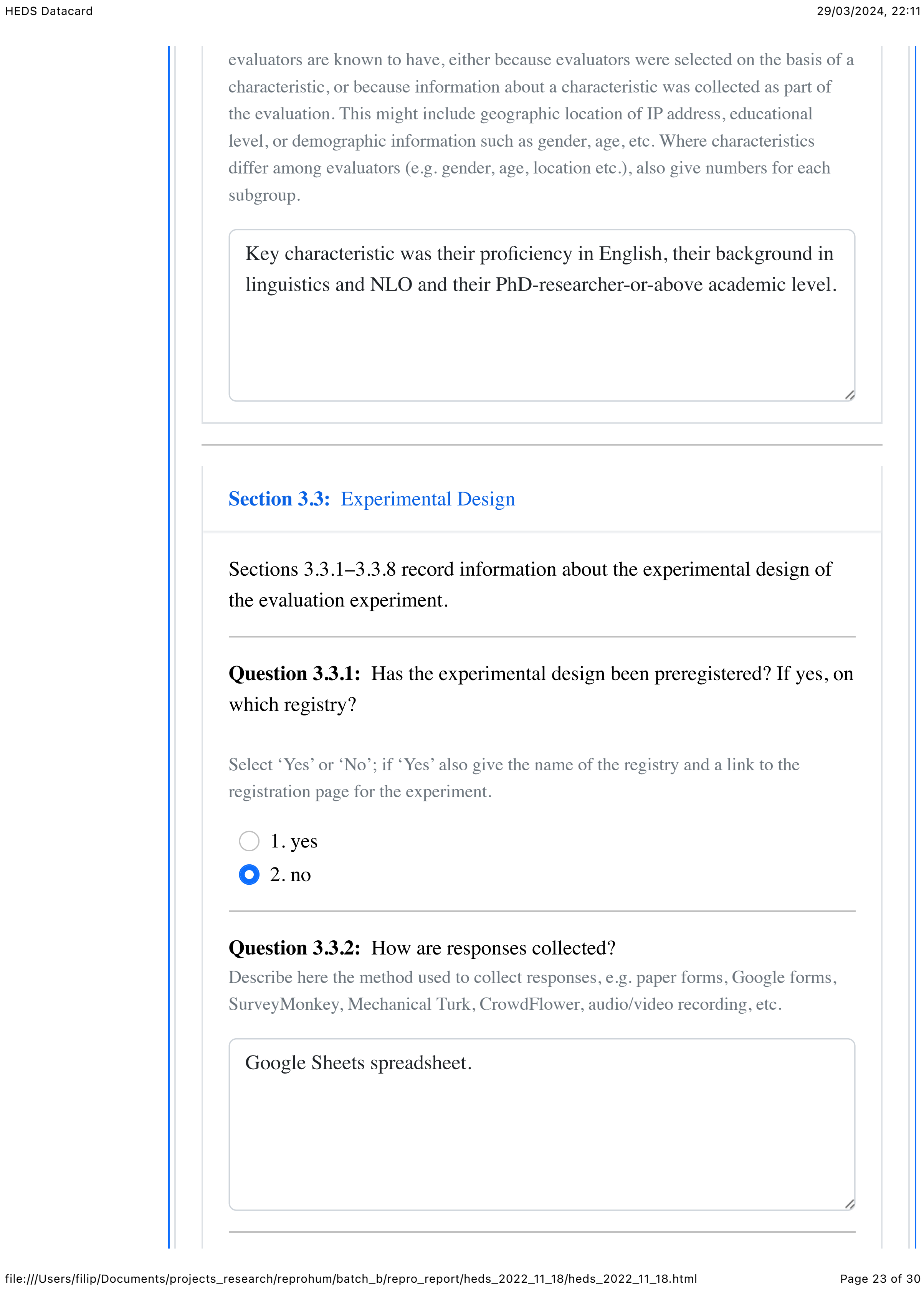
Question 3.2.4: What training and/or practice are evaluators given before starting on the evaluation itself?

Use this space to describe any training evaluators were given as part of the experiment to prepare them for the evaluation task, including any practice evaluations they did. This includes any introductory explanations they're given, e.g. on the start page of an online evaluation tool.

We sent them brief instructions via email and a definition of the redundancy quality criterion. Annotator training and guidelines were minimal to mirror the setting in the original study.

Question 3.2.5: What other characteristics do the evaluators have? Known either because these were qualifying criteria, or from information gathered as part of the evaluation.

Use this space to list any characteristics not covered in previous questions that the



evaluators are known to have, either because evaluators were selected on the basis of a characteristic, or because information about a characteristic was collected as part of the evaluation. This might include geographic location of IP address, educational level, or demographic information such as gender, age, etc. Where characteristics differ among evaluators (e.g. gender, age, location etc.), also give numbers for each subgroup.

Key characteristic was their proficiency in English, their background in linguistics and NLO and their PhD-researcher-or-above academic level.

Section 3.3: Experimental Design

Sections 3.3.1–3.3.8 record information about the experimental design of the evaluation experiment.

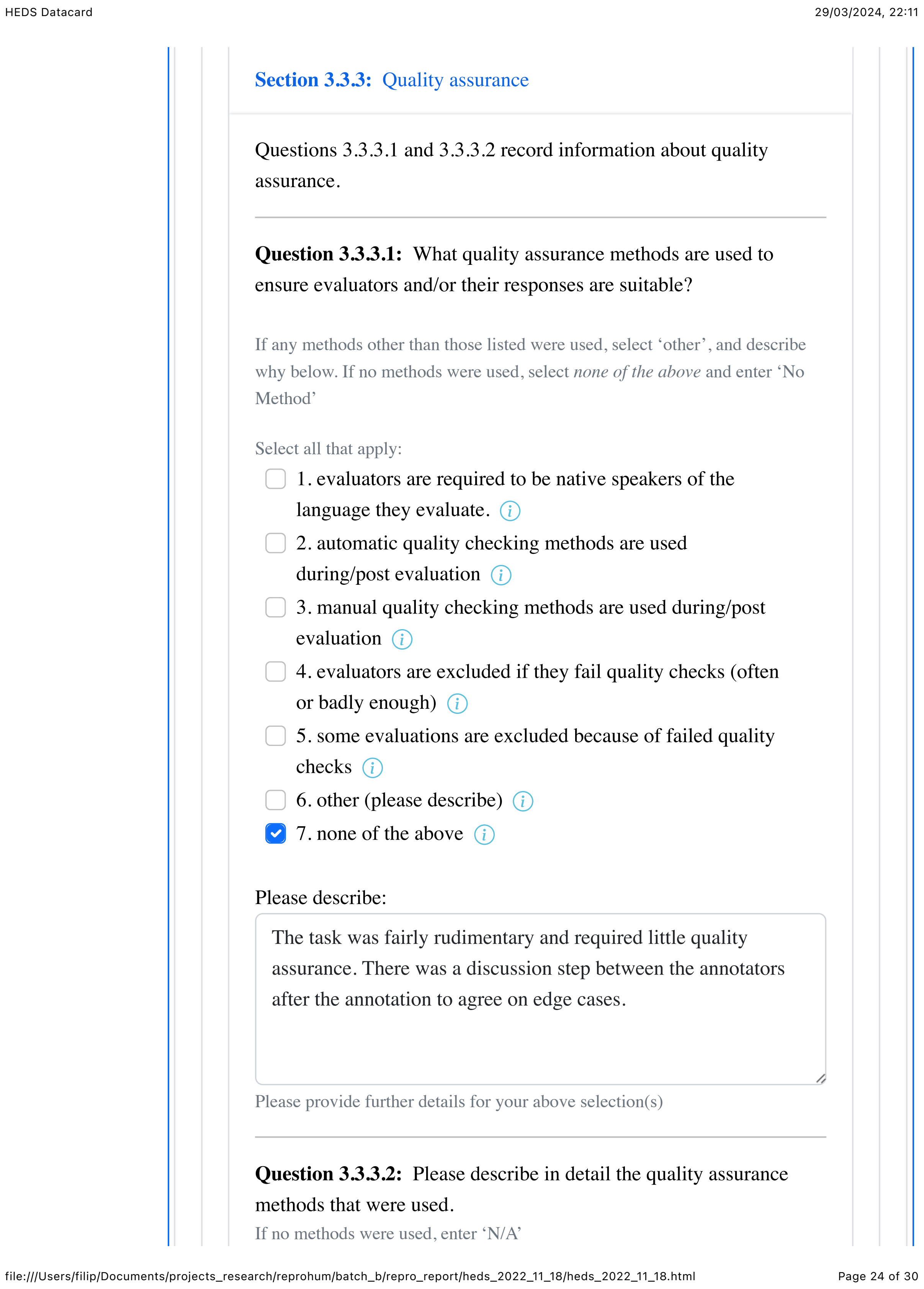
Question 3.3.1: Has the experimental design been preregistered? If yes, on which registry?

Select 'Yes' or 'No'; if 'Yes' also give the name of the registry and a link to the registration page for the experiment.

○ 1. yes
● 2. no

Question 3.3.2: How are responses collected?
Describe here the method used to collect responses, e.g. paper forms, Google forms, SurveyMonkey, Mechanical Turk, CrowdFlower, audio/video recording, etc.

Google Sheets spreadsheet.



Section 3.3.3: Quality assurance

Questions 3.3.3.1 and 3.3.3.2 record information about quality assurance.

Question 3.3.3.1: What quality assurance methods are used to ensure evaluators and/or their responses are suitable?

If any methods other than those listed were used, select ‘other’, and describe why below. If no methods were used, select *none of the above* and enter ‘No Method’

Select all that apply:

☐ 1. evaluators are required to be native speakers of the language they evaluate. ⓘ

☐ 2. automatic quality checking methods are used during/post evaluation ⓘ

☐ 3. manual quality checking methods are used during/post evaluation ⓘ

☐ 4. evaluators are excluded if they fail quality checks (often or badly enough) ⓘ

☐ 5. some evaluations are excluded because of failed quality checks ⓘ

☐ 6. other (please describe) ⓘ

☑ 7. none of the above ⓘ

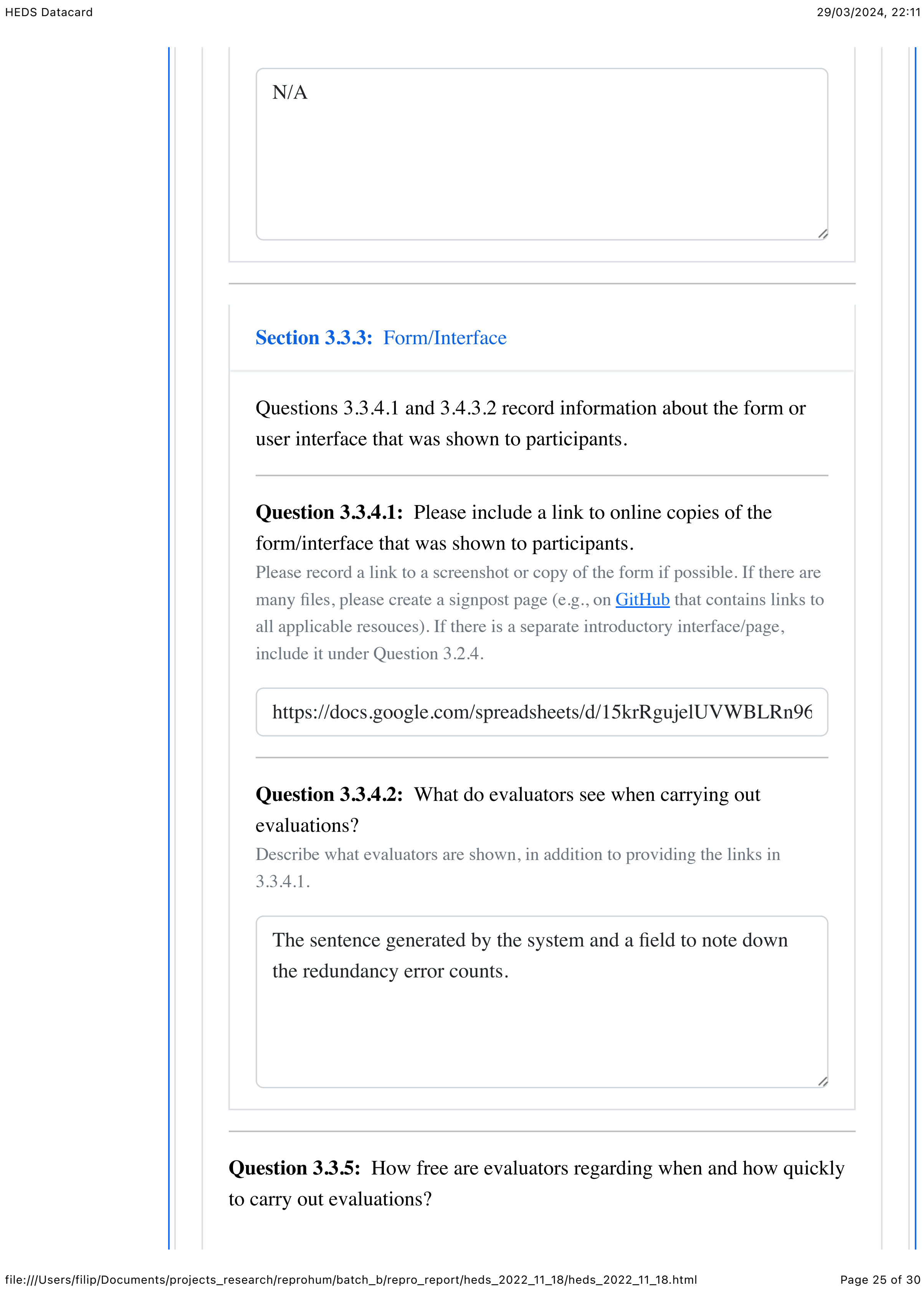
Please describe:

The task was fairly rudimentary and required little quality assurance. There was a discussion step between the annotators after the annotation to agree on edge cases.

Please provide further details for your above selection(s)

Question 3.3.3.2: Please describe in detail the quality assurance methods that were used.

If no methods were used, enter ‘N/A’



N/A

Section 3.3.3: Form/Interface

Questions 3.3.4.1 and 3.4.3.2 record information about the form or user interface that was shown to participants.

Question 3.3.4.1: Please include a link to online copies of the form/interface that was shown to participants.

Please record a link to a screenshot or copy of the form if possible. If there are many files, please create a signpost page (e.g., on GitHub that contains links to all applicable resouces). If there is a separate introductory interface/page, include it under Question 3.2.4.

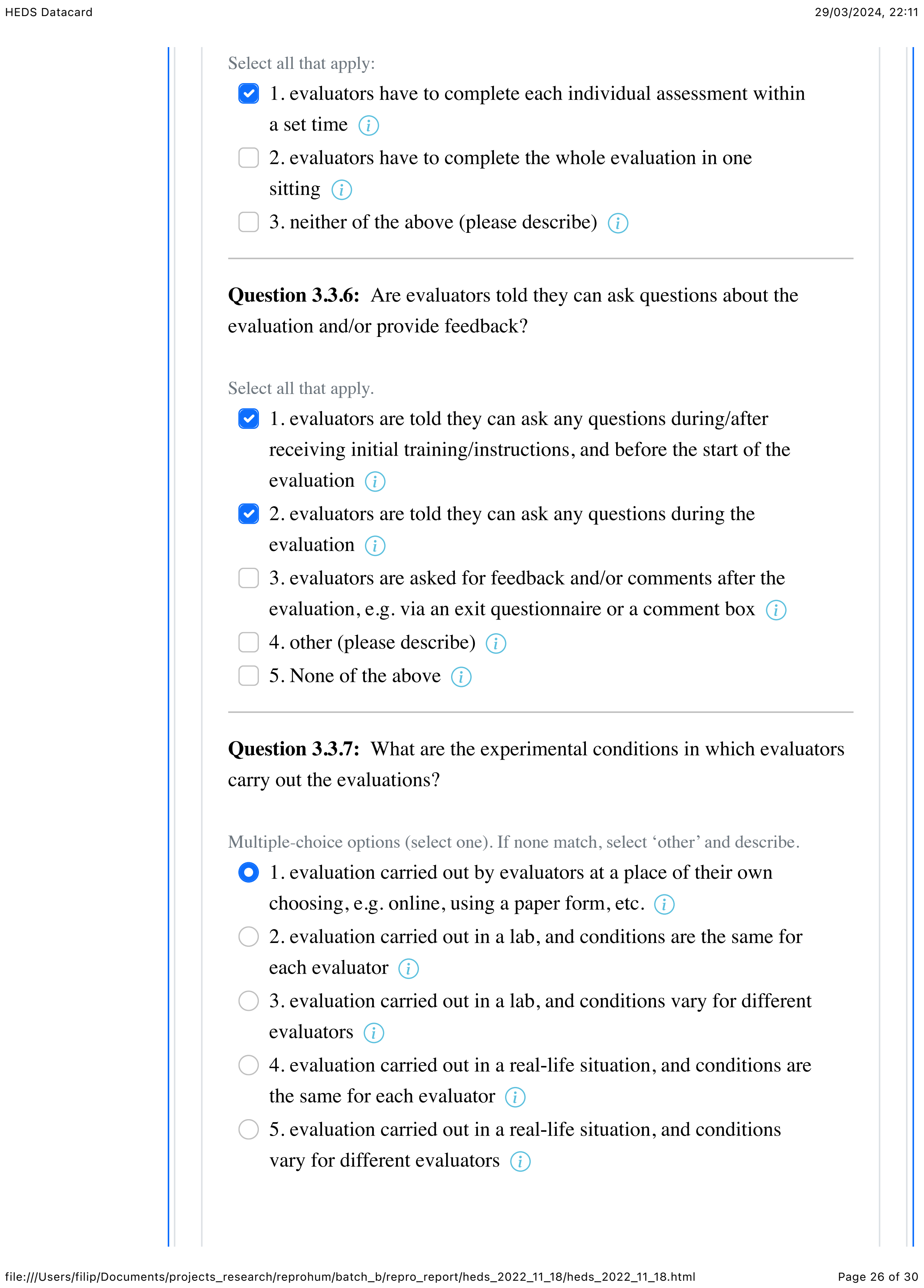
https://docs.google.com/spreadsheets/d/15krRgujelUVWBLRn96

Question 3.3.4.2: What do evaluators see when carrying out evaluations?

Describe what evaluators are shown, in addition to providing the links in 3.3.4.1.

The sentence generated by the system and a field to note down the redundancy error counts.

Question 3.3.5: How free are evaluators regarding when and how quickly to carry out evaluations?



HEDS Datacard 29/03/2024, 22:11

Select all that apply:

☑ 1. evaluators have to complete each individual assessment within a set time

□ 2. evaluators have to complete the whole evaluation in one sitting

□ 3. neither of the above (please describe)

Question 3.3.6: Are evaluators told they can ask questions about the evaluation and/or provide feedback?

Select all that apply.

☑ 1. evaluators are told they can ask any questions during/after receiving initial training/instructions, and before the start of the evaluation

☑ 2. evaluators are told they can ask any questions during the evaluation

□ 3. evaluators are asked for feedback and/or comments after the evaluation, e.g. via an exit questionnaire or a comment box

□ 4. other (please describe)

□ 5. None of the above

Question 3.3.7: What are the experimental conditions in which evaluators carry out the evaluations?

Multiple-choice options (select one). If none match, select ‘other’ and describe.

● 1. evaluation carried out by evaluators at a place of their own choosing, e.g. online, using a paper form, etc.

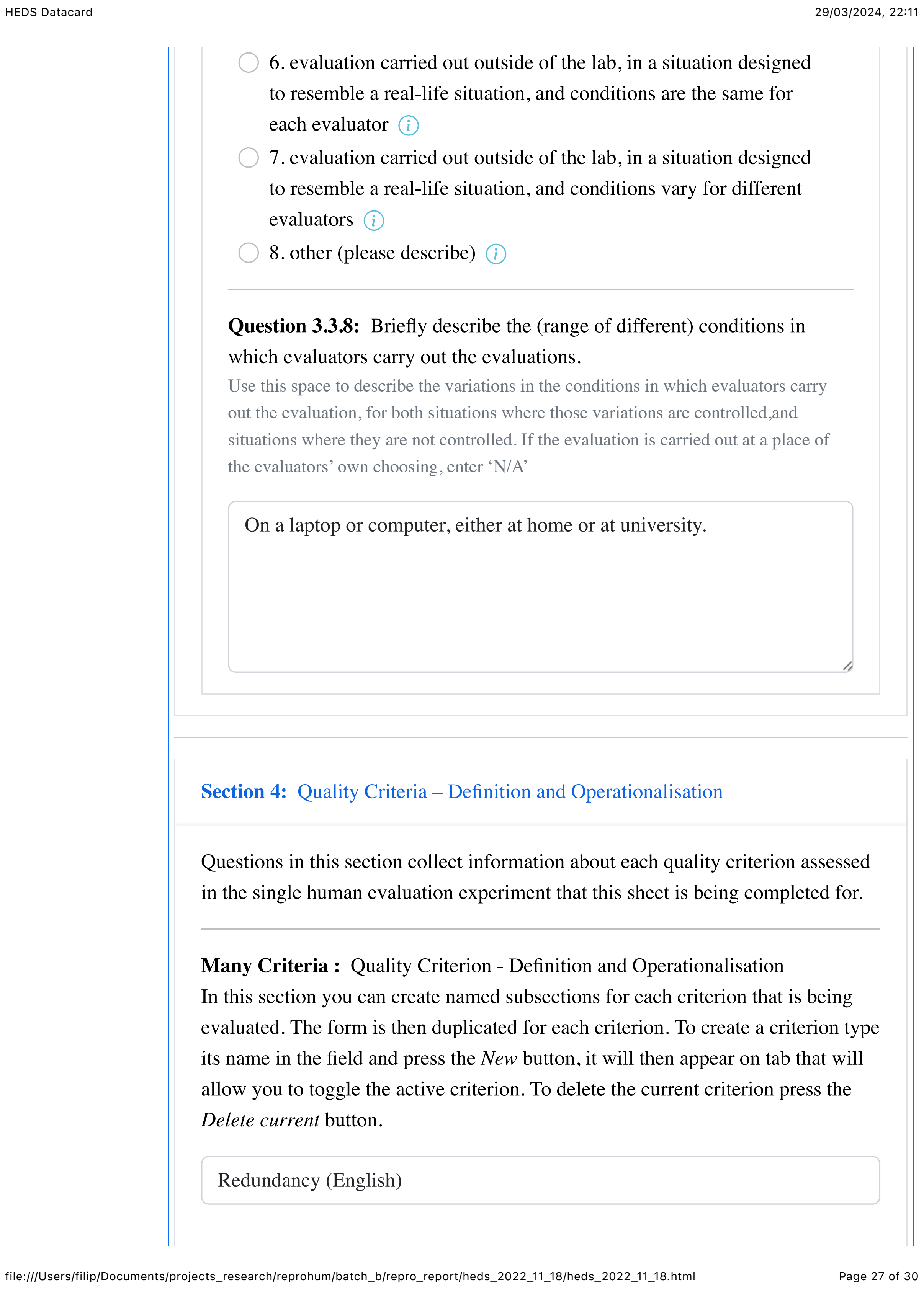
○ 2. evaluation carried out in a lab, and conditions are the same for each evaluator

○ 3. evaluation carried out in a lab, and conditions vary for different evaluators

○ 4. evaluation carried out in a real-life situation, and conditions are the same for each evaluator

○ 5. evaluation carried out in a real-life situation, and conditions vary for different evaluators

file:///Users/filip/Documents/projects_research/reprohum/batch_b/repro_report/heds_2022_11_18/heds_2022_11_18.html Page 26 of 30



○ 6. evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions are the same for each evaluator ⓘ

○ 7. evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions vary for different evaluators ⓘ

○ 8. other (please describe) ⓘ

Question 3.3.8: Briefly describe the (range of different) conditions in which evaluators carry out the evaluations.

Use this space to describe the variations in the conditions in which evaluators carry out the evaluation, for both situations where those variations are controlled,and situations where they are not controlled. If the evaluation is carried out at a place of the evaluators' own choosing, enter 'N/A'

On a laptop or computer, either at home or at university.

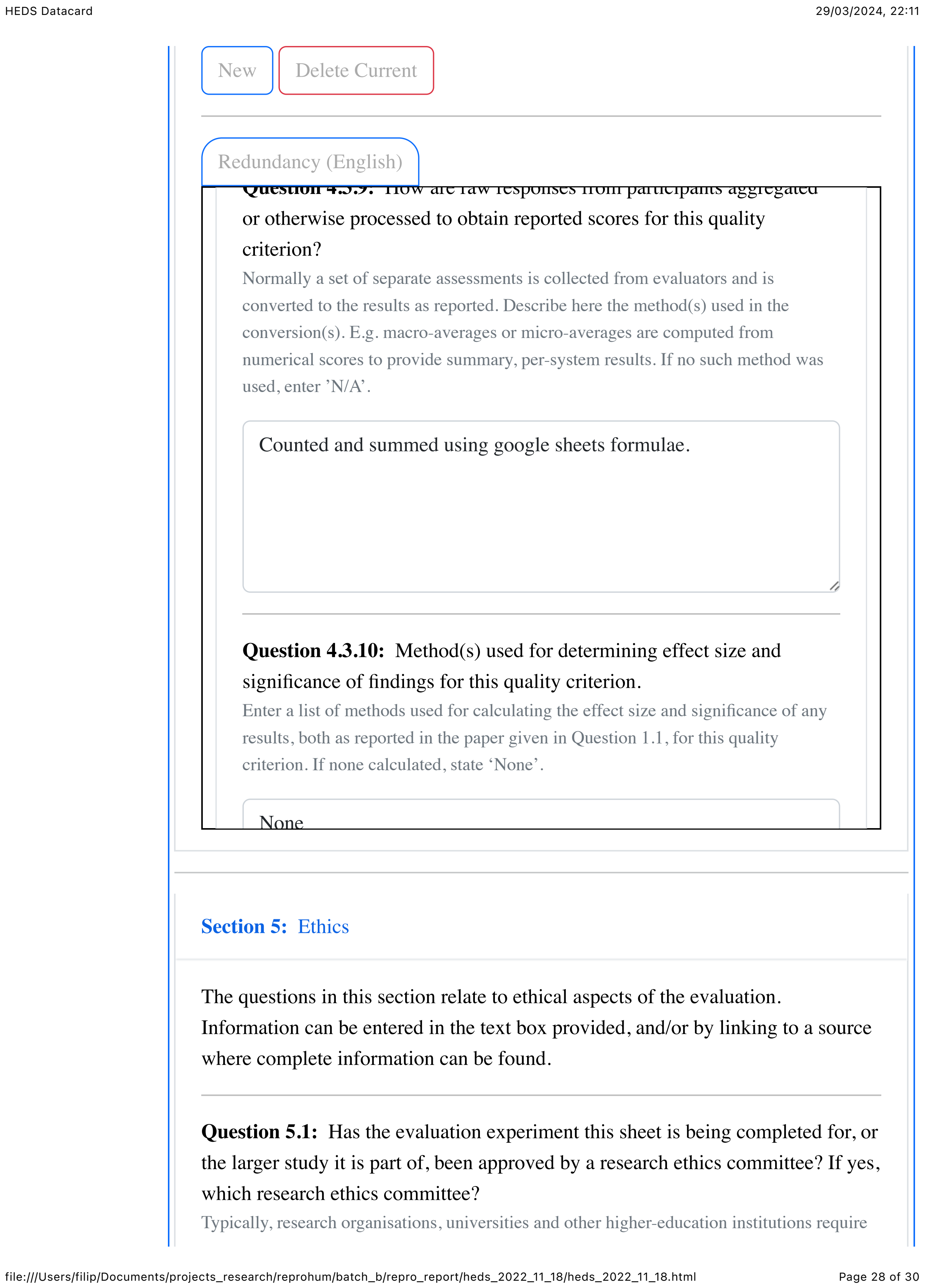
Section 4: Quality Criteria – Definition and Operationalisation

Questions in this section collect information about each quality criterion assessed in the single human evaluation experiment that this sheet is being completed for.

Many Criteria : Quality Criterion - Definition and Operationalisation

In this section you can create named subsections for each criterion that is being evaluated. The form is then duplicated for each criterion. To create a criterion type its name in the field and press the *New* button, it will then appear on tab that will allow you to toggle the active criterion. To delete the current criterion press the *Delete current* button.

Redundancy (English)



Question 4.3.9: How are raw responses from participants aggregated or otherwise processed to obtain reported scores for this quality criterion?

Normally a set of separate assessments is collected from evaluators and is converted to the results as reported. Describe here the method(s) used in the conversion(s). E.g. macro-averages or micro-averages are computed from numerical scores to provide summary, per-system results. If no such method was used, enter ’N/A’.

Counted and summed using google sheets formulae.

Question 4.3.10: Method(s) used for determining effect size and significance of findings for this quality criterion.

Enter a list of methods used for calculating the effect size and significance of any results, both as reported in the paper given in Question 1.1, for this quality criterion. If none calculated, state ‘None’.

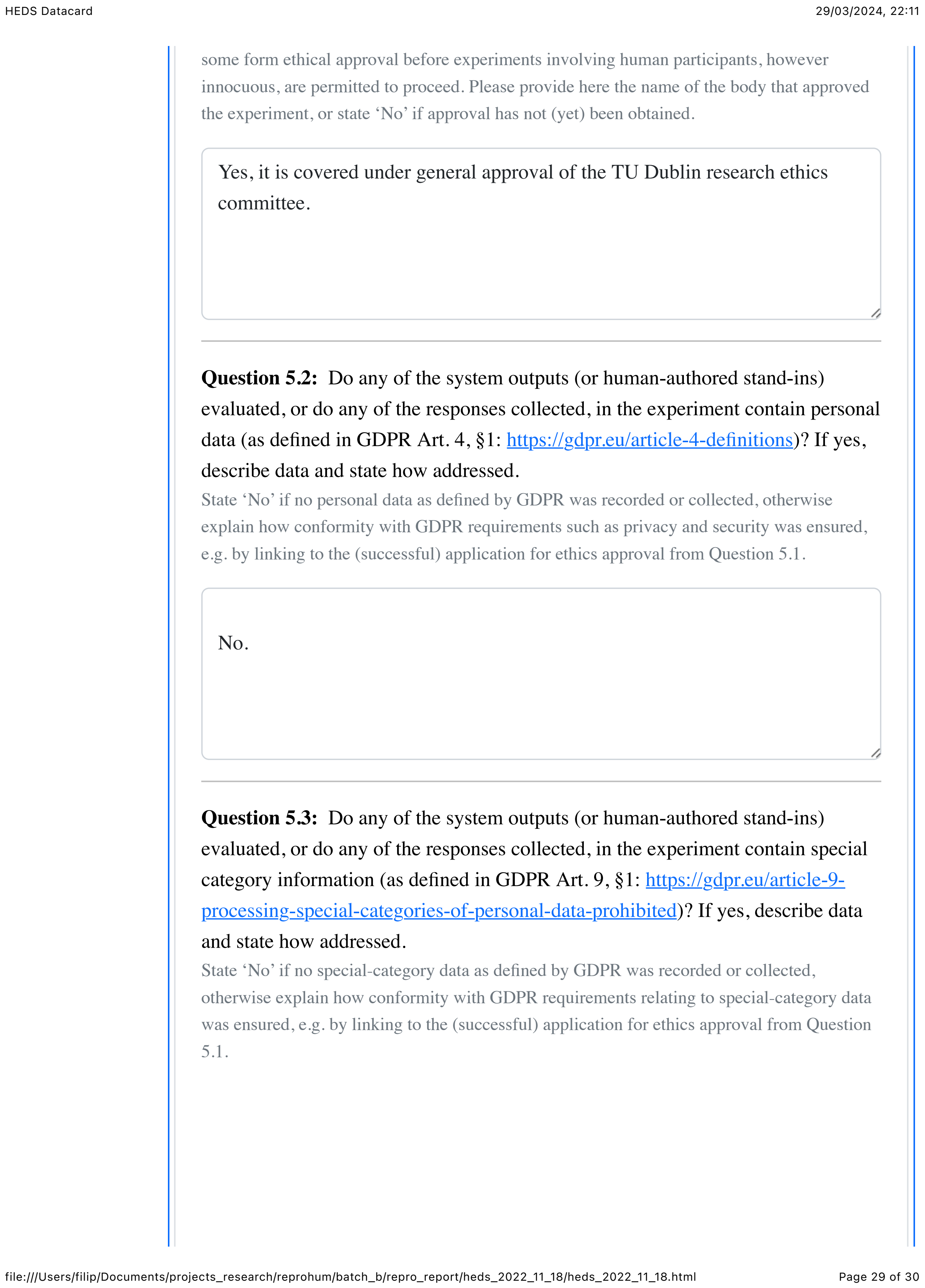
None

Section 5: Ethics

The questions in this section relate to ethical aspects of the evaluation. Information can be entered in the text box provided, and/or by linking to a source where complete information can be found.

Question 5.1: Has the evaluation experiment this sheet is being completed for, or the larger study it is part of, been approved by a research ethics committee? If yes, which research ethics committee?

Typically, research organisations, universities and other higher-education institutions require



some form ethical approval before experiments involving human participants, however innocuous, are permitted to proceed. Please provide here the name of the body that approved the experiment, or state 'No' if approval has not (yet) been obtained.

Yes, it is covered under general approval of the TU Dublin research ethics committee.

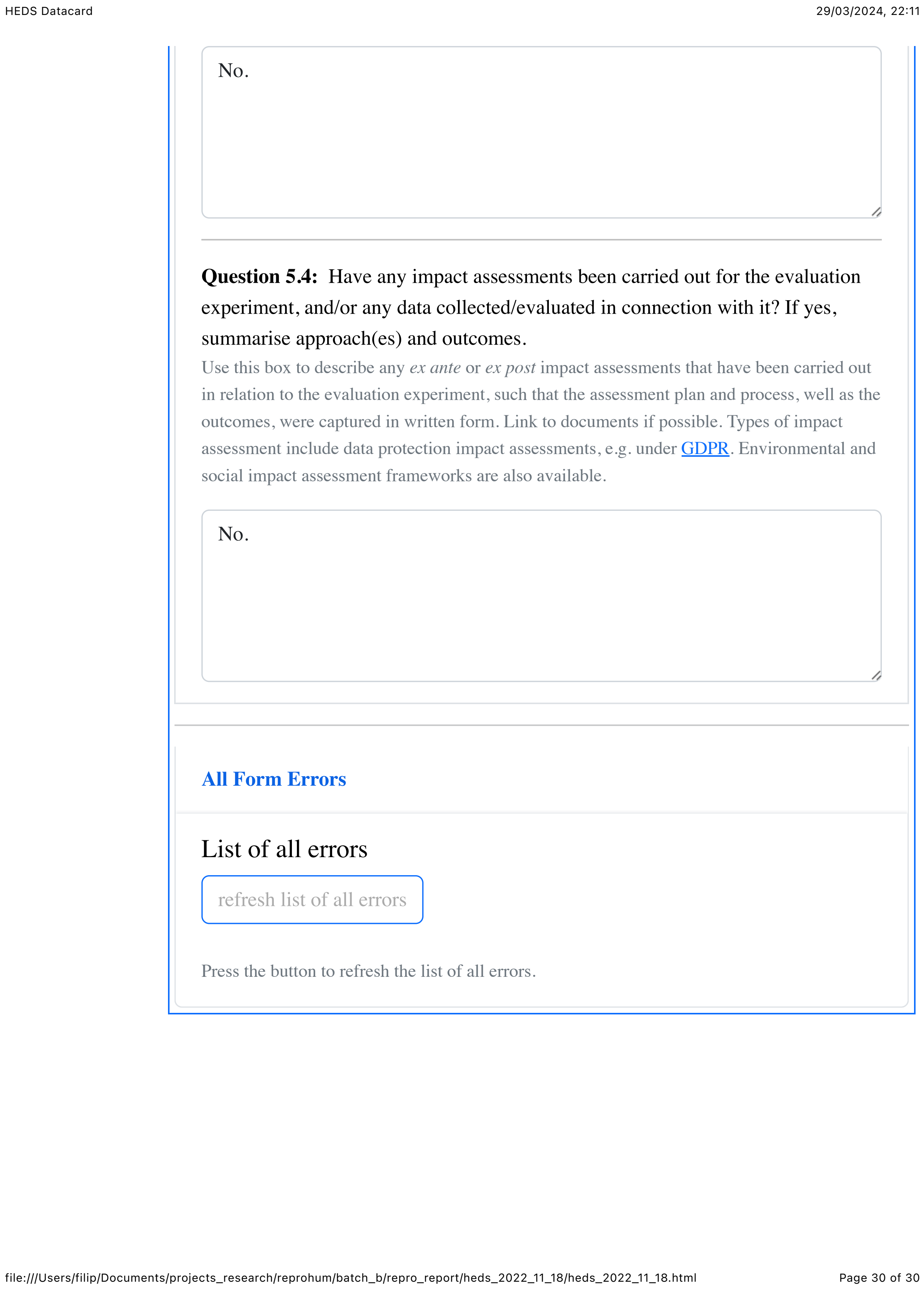
Question 5.2: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain personal data (as defined in GDPR Art. 4, §1: https://gdpr.eu/article-4-definitions)? If yes, describe data and state how addressed.

State 'No' if no personal data as defined by GDPR was recorded or collected, otherwise explain how conformity with GDPR requirements such as privacy and security was ensured, e.g. by linking to the (successful) application for ethics approval from Question 5.1.

No.

Question 5.3: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain special category information (as defined in GDPR Art. 9, §1: https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited)? If yes, describe data and state how addressed.

State 'No' if no special-category data as defined by GDPR was recorded or collected, otherwise explain how conformity with GDPR requirements relating to special-category data was ensured, e.g. by linking to the (successful) application for ethics approval from Question 5.1.



No.

Question 5.4: Have any impact assessments been carried out for the evaluation experiment, and/or any data collected/evaluated in connection with it? If yes, summarise approach(es) and outcomes.

Use this box to describe any *ex ante* or *ex post* impact assessments that have been carried out in relation to the evaluation experiment, such that the assessment plan and process, well as the outcomes, were captured in written form. Link to documents if possible. Types of impact assessment include data protection impact assessments, e.g. under GDPR. Environmental and social impact assessment frameworks are also available.

No.

All Form Errors

List of all errors

refresh list of all errors

Press the button to refresh the list of all errors.