PROTOCOL: Annotation by chemistry expert #2 for Addressing Unreliability Propagation in Scientific Digital Libraries

Roles

- one programmer
- two experts
- two modelers

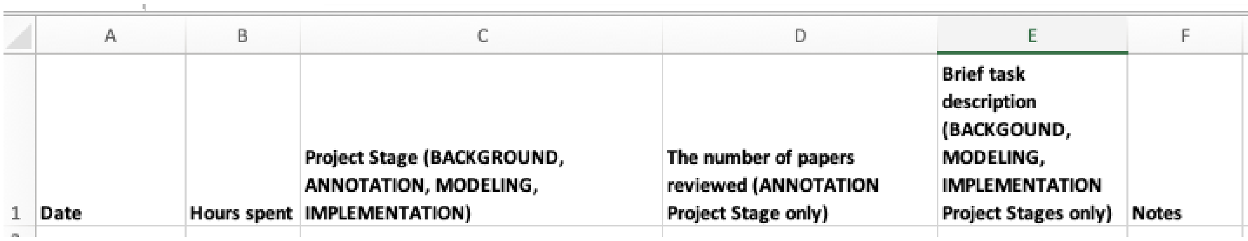
Protocol

Sample the citation contexts into a DOI-level dataset

1. The programmer uses NLP techniques to select a representative sample of n = ~80 citation contexts from citing publications of the WJH protocol. *We chose up to 80 citation contexts because this amount will require no more than 8 to 10 hours for an expert to process based on experience.* The corresponding full texts will also be needed. The programmer first selects the optimal number of clusters:
    a. Generate 2-10 clusters using BERTopic. *BERTopic uses the HDBSCAN clustering method by default, which aims to optimally identify the number of clusters.* To generate a specific number of clusters, use the KMeans clustering method instead.
    b. Select the best model (number of clusters) using coherence.
2. The programmer generates a sample of n citation contexts:
    a. Ideal scenario: we have c clusters and each cluster has at least n/c citation contexts. Randomly select n/c citation contexts for each cluster.
    b. If some clusters have less than n/c citation contexts, select all citation contexts from those clusters. After this step, we need to select x remaining citation contexts from the bigger clusters. Randomly select the x citation contexts using stratified random sampling from the bigger clusters.
3. Generate a list of DOIs for the 2nd expert's annotation work by deduplicating DOIs that appear from multiple citation contexts. Prepare an annotation spreadsheet.

Annotation and analysis of the DOI-level dataset

1. Orientation for the 2nd expert:
   a. Discuss the timetracking spreadsheet.[1]

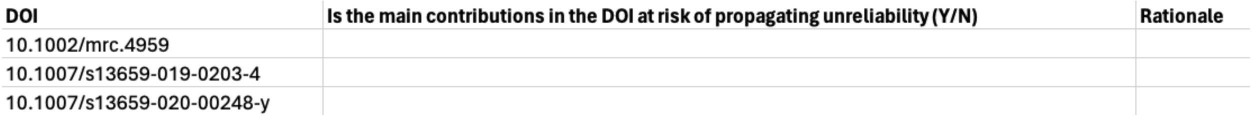


	A	B	C	D	E	F
1	**Date**	**Hours spent**	**Project Stage (BACKGROUND, ANNOTATION, MODELING, IMPLEMENTATION)**	**The number of papers reviewed (ANNOTATION Project Stage only)**	**Brief task description (BACKGOUND, MODELING, IMPLEMENTATION Project Stages only)**	**Notes**

The 2nd expert will record:

- Date
- Hours spent
- Project stage (BACKGROUND, ANNOTATION, MODELING)
- In the ANNOTATION Project Stage only: the number of papers reviewed (in that session)
- In the BACKGROUND and MODELING Project Stages only: a brief task description
- Notes (if any) may be recorded.

   b. Discuss the annotation spreadsheet.



DOI	**Is the main contributions in the DOI at risk of propagating unreliability (Y/N)**	**Rationale**
10.1002/mrc.4959		
10.1007/s13659-019-0203-4		
10.1007/s13659-020-00248-y		

Provided on the spreadsheet:
DOI - Digital Object Identifier

Also provided is a folder of PDFs. Filename of each PDF is its DOI. In each PDF the citation contexts of Willoughby et al. (2014) and its bibliography entry have been highlighted.

The 2nd expert will record:

[1] NOTE: The 2nd modeler will use a similar sheet to track MODELING and IMPLEMENTATION time only. The programmer will use a similar sheet to track IMPLEMENTATION time only.

- Is the main contribution in the DOI at risk of propagating unreliability (Y/N):
  Y - The DOI is at risk of propagating unreliability
  N - The DOI is NOT at risk of propagating unreliability
- Rationale - Justification of the Y/N decision

2. BACKGROUND: The 2nd expert reads and understands (1) Willoughby et al. (2014) to understand the paper; and (2) Bhandari Neupane et al. (2019) and Willoughby et al. (2020) to understand the nature of the code glitch and its correction.
   *2nd expert: TIME TRACKING for Stage **BACKGROUND***

3. ANNOTATION: The 2nd expert makes decisions about whether the citing publication's main contribution is at risk due to the code glitch and records the decisions on **2nd_expert_annotation_spreadsheet.csv**

   DOI - Digital Object Identifier

   Is the main contribution in the DOI at risk of propagating unreliability (Y/N) - A decision of whether the main contributions in the DOI at risk of propagating unreliability

   Rationale - The reason to justify the decision

   *2nd expert: TIME TRACKING for Stage **ANNOTATION***

4. 1st modeler checks the inter-annotator agreement between the 1st expert and the 2nd expert at the DOI level and calculates the percentage disagreement.
   a. 1st expert reviews the discrepancies and characterizes them.
   b. If more than 10% disagreement, the two modelers & 1st expert decide whether a 3rd expert is needed.

5. MODELING: The 2nd modeler helps the 2nd expert model their decision-making process into a decision tree.
   *2nd expert: TIME TRACKING for Stage **MODELING***
   *2nd modeler: TIME TRACKING for Stage **MODELING***
   a. The 2nd modeler compares the decision trees.

6. IMPLEMENTATION: The programmer and 2nd modeler work with the 2nd expert to implement the 2nd decision tree computationally.
2nd expert: TIME TRACKING for Stage ***IMPLEMENTATION***
2nd modeler: TIME TRACKING for Stage ***IMPLEMENTATION***
programmer: TIME TRACKING for Stage ***IMPLEMENTATION***

7. The programmer runs the system using the 2nd decision tree and obtains machine decisions on the full list of 284 publications, and a list of differences with the 1st expert's manual annotation.

Evaluation metrics

1. Percentage disagreement between the 1st expert and the 2nd expert (step 4a)
2. Accuracy of the machine results compared to the 1st expert’s annotation (step 7)
    a. Number that agree / total number of instances
3. Time requirements:
    a. For BACKGROUND, ANNOTATION, MODELING, and IMPLEMENTATION, the 2nd expert needs to track the time they spend.
    b. For MODELING and IMPLEMENTATION, the 2nd modeler needs to track the time they spend in collaboration with the expert.
    c. For IMPLEMENTATION steps, the programmer tracks the time they spend.

Bibliography

Bhandari Neupane, J., Neupane, R. P., Luo, Y., Yoshida, W. Y., Sun, R., & Williams, P. G. (2019). Characterization of Leptazolines A–D, polar oxazolines from the cyanobacterium Leptolyngbya sp., reveals a glitch with the “Willoughby–Hoye” scripts for calculating NMR chemical shifts. Organic Letters, 21(20), 8449–8453. https://doi.org/10.1021/acs.orglett.9b03216

Willoughby, P. H., Jansma, M. J., & Hoye, T. R. (2014). A guide to small-molecule structure assignment through computation of (^{1}H and ^{13}C) NMR chemical shifts. Nature Protocols, 9(3), Article 3. https://doi.org/10.1038/nprot.2014.042

Willoughby, P.H., Jansma, M.J. & Hoye, T.R. (2020). Addendum: A guide to small-molecule structure assignment through computation of (^{1}H and ^{13}C) NMR chemical shifts. Nat Protocols, 15, 2277. https://doi.org/10.1038/s41596-020-0293-9