Qualitative Sensibilities for Data Science Research Pipeline
(working draft)

Sarah Dreier (University of Washington) and Emily Gade (Emory University)

January 2021

PRE-CODING RESEARCH DEVELOPMENT	
Research design	Develop your research question and execution plan.
Brain dump to record researcher motivations and preconceptions	What are your pre-existing assumptions, motivations, and expectations?
Situational mapping to record researcher's relationship to the topic and results	What stakes do you have in the results? How is your perspective shaped by your academic discipline (e.g., political science, computer science, history), your own political identity (e.g., as a citizen of a given country, a member of a racial group, or a member of a political party), your lived experiences, and/or your career goals (e.g., achieving tenure or maximizing industry advantages)?
Toolbox critique to identify and address strengths and weakness	**Data**: Examine the value and limitations to your data and planned methodological approach. Who generated the data and what material, information, or perspectives might it exclude or under-represent? *Ex: We researched the UK National Archives' curation process, scope conditions, and redaction policies, and we considered strategies for identifying and incorporating (via additional sources) perspectives might be excluded or under-represented.* **Methods:** Why did you select your planned methodologies? What do you gain from using those methods? What do you lose or overlook from using these methods, and how might you mitigate those short-comings? How might your methods introduce unexpected challenges when applied to your data?
Case-study knowledge to ensure detailed, contextual accuracy	Review academic research, historical records, news coverage, and other sources that provide you a clear understanding of the relevant events, actors, arguments, policies, public debates, timeframe, and context. Read beyond your own disciplinary boundaries and seek to read material from a variety of actors' vantage points.
Revisit research design based on these steps	Review and update your research design (if/as appropriate) to accommodate the insights, possible concerns, and case study knowledge identified above.
ONTOLOGY AND PRELIMINARY CODING	
Develop codes and execute initial coding to accommodate abductive reasoning and establish conceptually and contextually meaningful coding approaches	• Batch code for broad themes that define the universe of data –OR– Review a small subset of data to familiarize yourself with the data. • Deductively develop a codebook. Make a list of the categories/concepts you expect to see (name, description, hypothetical example). Make notes about where you might expect to see overlap or blurred boundaries between categories. Include "miscellaneous" and "open relevant" codes to accommodate unexpected observations or relationships.

	• Conduct pilot coding among a subset of data. Keep a coding log that records judgment calls, questions or confusion, issues, and new (unanticipated) categories or concepts that emerge in the data. • Revisit codebook to update, shift, add, remove, merge, or separate categories or concepts as appropriate, based on your pilot coding. • Discuss your updated coding approach with your team or other colleagues to ensure the approach is reasonable and meaningful. • Team coding exercise. Have 2-3 members of your team (if possible) code the same few documents to identify and discuss coding disagreements and update your coding approach as appropriate.
CODE (AND/OR ANALYZE) DATA	
Evaluate inter-coder reliability to ensure coding replicability	Train coders based on your coding ontology, randomly select material for them to code, calculate Kappa scores to assess inter-coder agreement, and address any issues that arise as appropriate.
Code main dataset	Code the main set of your data that requires annotation (the whole corpus or a subset of training data). Adhere to your updated coding ontology as best as possible.
Maintain fieldnotes to facilitate interpretive, abductive, and reflexive sensibilities.	Maintain a daily fieldnote coding log. Record the following: • Any coding judgments and issues with the coding ontology. If necessary, **abductively** and modestly update coding approach (and document when the update occurred). • Observations about broad trends and connections between your concepts of interest. How do you understand or **interpret** those observations? These observations can be used as analytic evidence. • **Reflections** about how your own intuitions, assumptions, or perspective may be shaping your coding.
CONDUCT AND VALIDATE COMPUTATIONAL ANALYSIS	
Identify methods assumptions to avoid modeling errors	Carefully consider if/how your data differs from the forms of data for which your selected method was developed. Consider these deviations as appropriate. *Ex: Real-world text data introduces challenges which NLP practitioners may be unaccustomed to recognizing. Archives often contain full or partial duplication (e.g., multiple drafts of a statement) which require specific modeling attention (e.g., to keeping all duplicate text in the same training data split).*
Conduct computational analysis	
Validate results to catch errors and aid interpretation	Qualitatively examine model outputs to: identify modeling errors, ensure models are capturing the intended concepts, and gain increased familiarity with your data and results. *Do these results seem reasonable?*
Revisit research development steps to inform and situate conclusions	How might your personal situation as a researcher and your prior knowledge be shaping your model outputs or how you interpret those results? Do your conclusions make sense, given your case study knowledge? What knowledge or ideas did you presume earlier that you should dispel, revisit, or complicate? How will your conclusions impact various stakeholders or vulnerable populations? How well would your conclusions generalize to other contexts, and/or are these results shaped by the idiosyncrasies of your case study?