Evaluating alternative data fitness for use

Melanie Santiago

Industry Pricing Branch Chief

Producer Price Index

U.S. Bureau of Labor Statistics

38th Voorburg Group Meeting

October 2023

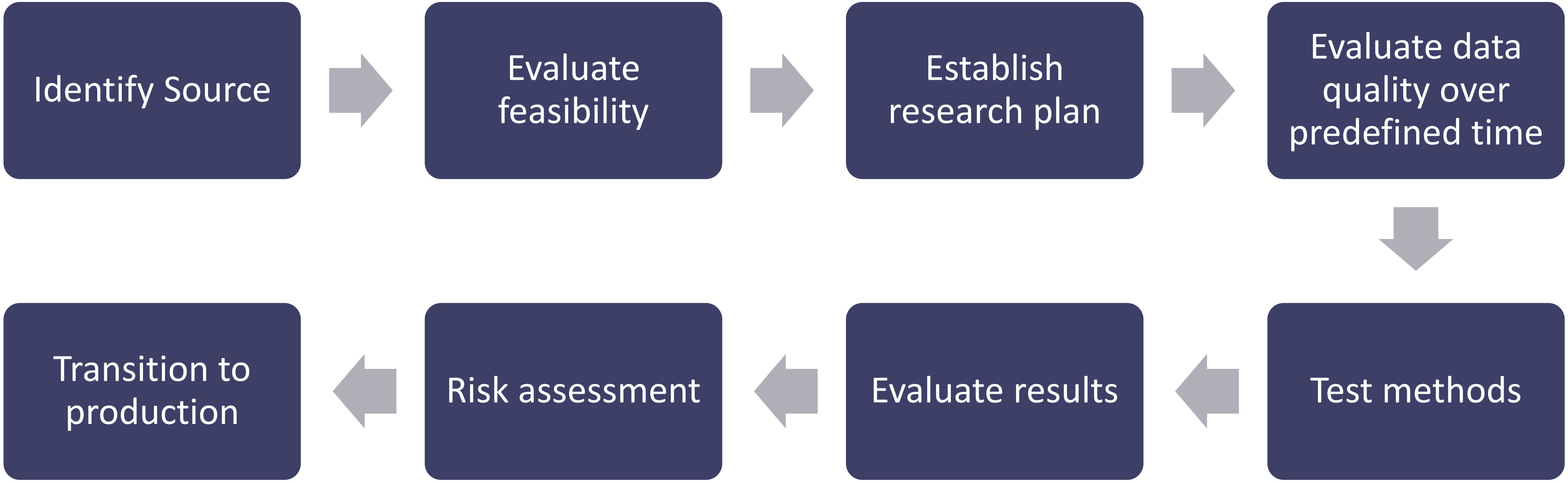
Background

- *Guidelines for Incorporating Alternative Data Sources in Official Statistics*
  - Presented at the 37th Voorburg Group meeting in 2022
  - Fitness for use questionnaire
  - Requested volunteers to test the questionnaire
- U.S. BLS evaluated two potential data sources and compared the questionnaire to our own evaluation process

Alternative data in the U.S. PPI

- Create sample frames
- Benchmark sample
- Supplement collected data to support hedonic modeling
- Replace/supplement current data collection methods
- Validation of survey data

General steps for alternative data projects



Identify Source

Evaluate feasibility

Establish research plan

Evaluate data quality over predefined time

Test methods

Evaluate results

Risk assessment

Transition to production

Overall impressions

- Questionnaire is very detailed and thorough
- Accounts for a wide variety of circumstances
- Included methodological considerations as well as practical implementation questions
- Not all questions will be relevant for every alternative data project
- May be daunting for those filling out the questionnaire

1. Specify needs

- Clear concise questions on the intended goals, costs, and timeliness
- 1.d *Does the data set exhibit the characteristics of an administrative data set? Or an alternative data set?*
  - U.S. defines administrative data as a subset of alternative data
  - May want to include definitions of terms within the questionnaire
  - Unclear what this question was asking

2. Design

- Clear, concise questions on the coverage of the data set and whether it adequately covers the target population
- Essential information to assess early in the process
- Aligns with questions on the US BLS scorecard

3. Build

- 3.a *What new components may be needed…? (e.g. data acquisition channel, data processing component, machine learning model evaluation, dissemination component).*
- These are additional costs not included in the price of a dataset.
  - Will these be one-time costs, or will there be maintenance?
  - Can these new systems/processes be used for other datasets?

4. Collect

- Collection and consistency questions were difficult to answer. This data is often not provided by vendors.
- 4.b *Is the data available at the level of granularity that is required to fulfill the needs of your statistical program?*
  - May be useful to ask what is and what is not available
  - Multiple dimensions of granularity (customers, quantities, configurations, geography, etc.)
  - Are there any important price determining characteristics missing?

5. Process

- Response errors and bias questions difficult to answer as data providers not always willing to provide this information
- 5.d *“…Does the NSI have the ability to maintain the independence of their statistical outputs with respect to the objective of the data provider or the originally intended use of the data?”*
    - This is important enough to be addressed as a stand-alone question.

6. Analysis

- 6.a *Are there obligations to the data provider or the constituent target population on the dissemination of data derived from the alternative data source? Do specific disclosure control measures need to be put in place?*
  - Terms of service and purchase agreements can impose many restrictions
  - What is meant by "obligations to … the constituent target population"?

7. Disseminate

- 7.a *Will the final data products replace existing statistics or be new to the NSI?*
- 7.b *Will they be "official" or "experimental" statistics?*
- Dissemination is at the end of the usual NSI processes
- Goals should be set early in the process
- Necessary to inform responses to parts 1 through 6

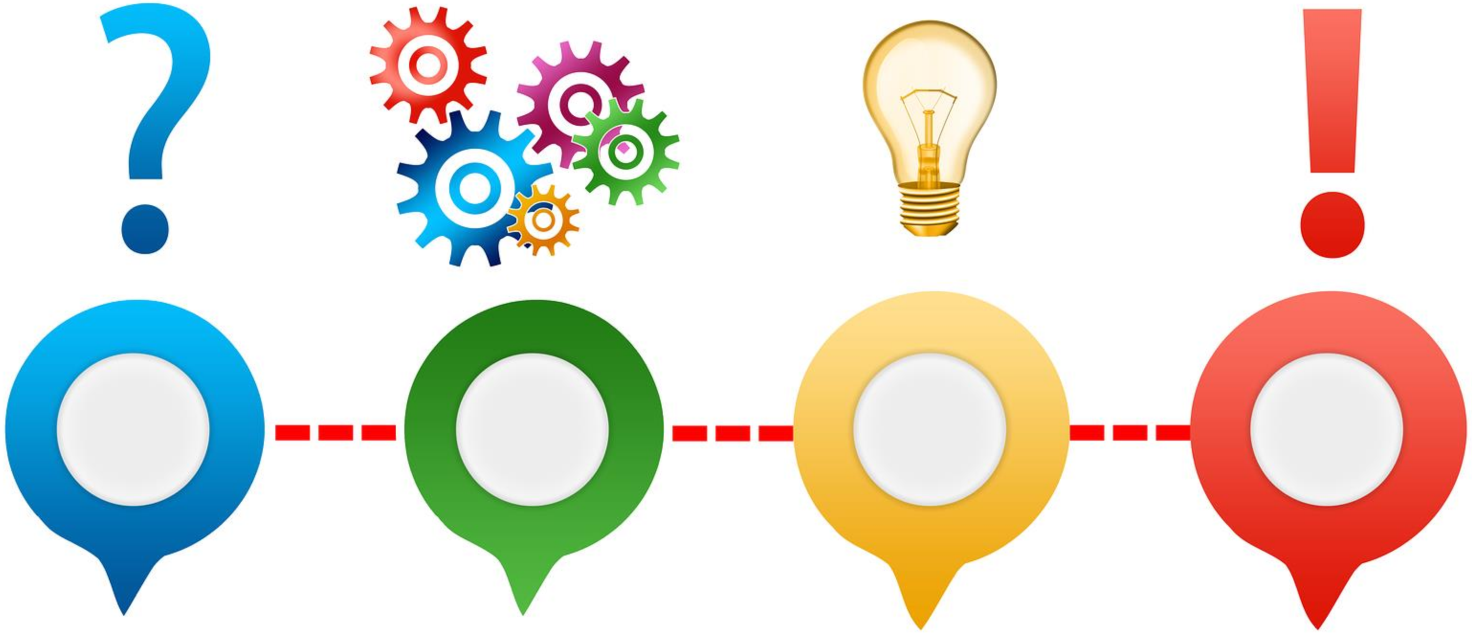
8. Evaluate

- *8. Each section of this questionnaire provides an opportunity to evaluate the statistical process as well as questions on data ethics. The NSI should review this questionnaire and record their reflections at various intervals…to ensure that expectations are realized and/or re-evaluated.*
    - This is more an introduction to how to use the questionnaire and would be more useful at the beginning

Final thoughts

- BLS splits this process into two steps:
  - Preliminary checklist to evaluate feasibility
  - In depth research to assess the dataset and develop an implementation plan
- The full questionnaire should only be completed if a preliminary analysis indicates the dataset may be viable
- BLS will be updating our evaluation procedures to incorporate some questions and scenarios from the questionnaire

Questions or Comments?



Contact Information

Melanie Santiago
Industry Pricing Branch Chief
Producer Price Index
www.bls.gov/ppi
202-691-7844
Santiago.Melanie@bls.gov