A practical guide to (successfully) collect and process images through online surveys

BigSurv 2023 Conference. Quito.

PATRICIA A. IGLESIAS | Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra

CARLOS OCHOA | Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra
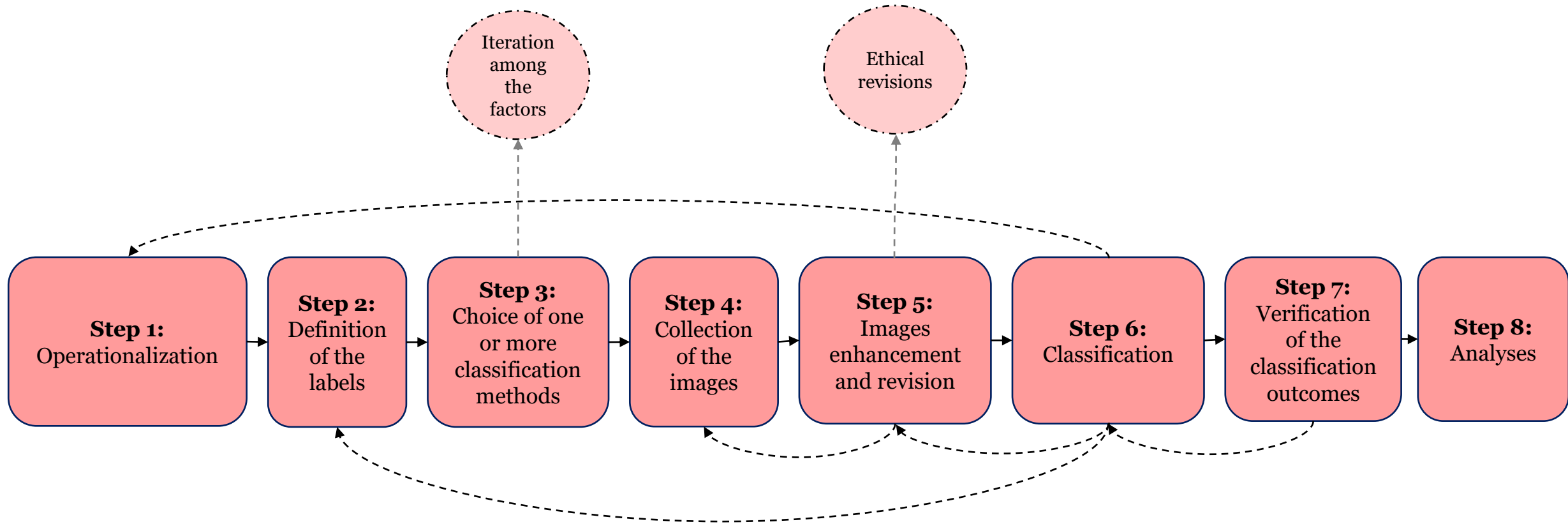
MELANIE REVILLA | Institut Barcelona d'Estudis Internacionals

Introduction

- Images as a new data source collected through surveys have gained relevance during the last years.

- They have the potential to **decrease the respondents' burden, increase the data accuracy and quality, and provide new insights** compared to conventional survey questions (Revilla, 2022).

- However, little has been said on **how to implement the collection of images in the frame of web surveys**, and on the main considerations that researchers should keep in mind when doing so.

- **We provide guidelines that practitioners can follow** before, during, and after the images' collection, in order to get the best of this type of data.

Our proposal to plan, collect, and analyze images: An eight-step process

The eight-step process to plan, collect, and analyze images in the frame of web surveys

web data opp

Iteration among the factors

Ethical revisions

Step 1: Operationalization

Step 2: Definition of the labels

Step 3: Choice of one or more classification methods

Step 4: Collection of the images

Step 5: Images enhancement and revision

Step 6: Classification

Step 7: Verification of the classification outcomes

Step 8: Analyses

This process is not necessarily linear

Step 1: Operationalization

(1)

Items → the information of interest present in the image.

(2) Objects

(3)

(4)

(5)

(6)

(7)

(8)

Category:
dog, blanket, floor

Attributes:
white, fabric, furry

Step 1: Operationalization

1 **Items →** the information of interest present in the image.

2

Objects

3

4 Scenes

5

6

7

8

Photo by Davey Gravy in Unsplash

Scene:
Friends meeting, celebration, dinner.

Step 1: Operationalization

1

Items → the information of interest present in the image.

2 Objects

3

4 Scenes

5

Text

6

7

8

```
                                    EUR
Banana                              1,44  A
     1,148 kg x 1,25 EUR/kg
Freshona/Espinacas                  1,15  A
Vemondo/Bebida soja 0%              1,60  B
     2 x 0,80
Vemondo/Tofu ecológico              0,95  B
Mandarina Ebre                      2,79  A
Dentalux/Crema dental               0,95  C
Chef Select/Trio de humm            2,19  B
Edulis/Ensalada dúo                 1,15  A
Alesto/Mezcla frutos sec            1,89  B
Floralys/Servill 2capas             0,95  C
Favorina/Huevos chocolat            1,49  B
Champiñón                           0,65  A
Huevos L suelo                      1,79  A
Floralys/Papel higiénico            2,55  C
                            ------------
Total                        21,54
                            ============
```

Focus:
Price per product

Step 1: Operationalization

1

Items → the information of interest present in the image.

2 Objects

3

4 Scenes

5

6 Text

7

8

```
                                    EUR
Banana                             1,44 A
     1,148 kg x 1,25 EUR/kg
Freshona/Espinacas                 1,15 A
Vemondo/Bebida soja 0%             1,60 B
     2 x 0,80
Vemondo/Tofu ecológico             0,95 B
Mandarina Ebre                     2,79 A
Dentalux/Crema dental              0,95 C
Chef Select/Trio de humm           2,19 B
Edulis/Ensalada dúo                1,15 A
Alesto/Mezcla frutos sec           1,89 B
Floralys/Servill 2capas            0,95 C
Favorina/Huevos chocolat           1,49 B
Champiñón                          0,65 A
Huevos L suelo                     1,79 A
Floralys/Papel higiénico           2,55 C
                            ------------
Total                          21,54
                            ============
```

Focus:
Products

Step 1: Operationalization

web data opp

1 **Items →** the information of interest present in the image.

2 Objects

3

Scenes

4

5

Text

6

7 More complex analysis

8

```
                              EUR
Banana                       1,44 A
    1,148 kg x 1,25 EUR/kg
Freshona/Espinacas           1,15 A
Vemondo/Bebida soja 0%       1,60 B
    2 x 0,80
Vemondo/Tofu ecológico       0,95 B
Mandarina Ebre               2,79 A
Dentalux/Crema dental        0,95 C
Chef Select/Trio de humm     2,19 B
Edulis/Ensalada dúo          1,15 A
Alesto/Mezcla frutos sec     1,89 B
Floralys/Servill 2capas      0,95 C
Favorina/Huevos chocolat     1,49 B
Champiñón                    0,65 A
Huevos L suelo               1,79 A
Floralys/Papel higiénico     2,55 C
                          ------------
Total                        21,54
                          ============
```

Type of document:
Receipt

Step 1: Operationalization

1

- Definition of the type(s) of item to be classified.
  - Categories and/or attributes

2

- Delimitation of the item(s) to be classified.
  - Some or all items in the image.

3

4

5

6

7

8
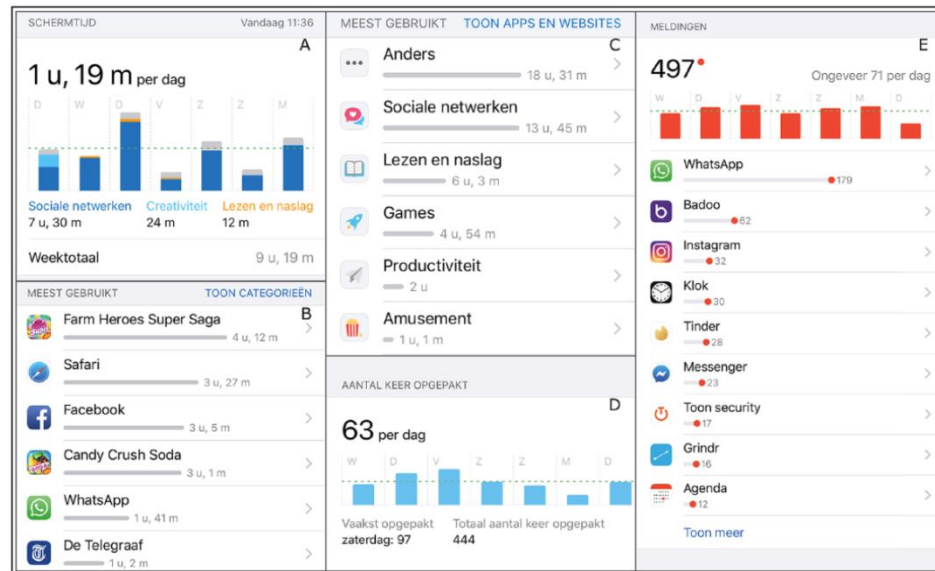
Ohme et al. (2020)

Step 2: Definition of the labels

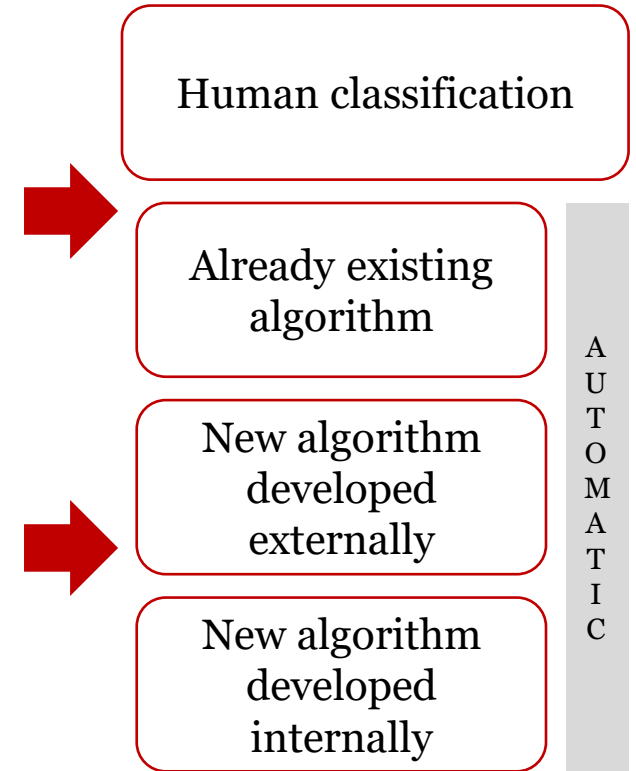- Think about the exact labels to be used → definition of the response scale.

Pataki et al. (2021)

Step 3: Choice of the classification method(s)

1
2
3
4
5
6
7
8

Factors for choosing the classification model

Features of the tasks	Resources	Overall data quality
Total number of images	Human resources	Accuracy
Total number of labels	Infrastructure	Consistency
Kind of labels	Budget	Data protection
Recurrence of the study	Time	Transparency
	Availability of images	

Human classification

Already existing algorithm

New algorithm developed externally

New algorithm developed internally

AUTOMATIC

• More than one method can be chosen

Step 3: Choice of the classification method(s)

1

2

3

4

5

6

7

8

- **Features of the tasks**

  - Total number of images
    - **If low (<500)**
    - If more

→ Human classification

Already existing algorithm

New algorithm developed externally

New algorithm developed internally

AUTOMATIC

Step 3: Choice of the classification method(s)

web data *opp*

1

2

3

4

5

6

7

8

- **Features of the tasks**

  - Total number of images
    - If low (<500)
    - **If more**

420 times faster and 221 times cheaper (Bosch et al., 2019)

Human classification

→ Already existing algorithm

→ New algorithm developed externally

→ New algorithm developed internally

A U T O M A T I C

Step 3: Choice of the classification method(s)

web data opp

1

2

3

4

5

6

7

8

- **Features of the tasks**

  - Total number of images
    - If low (<500)
    - If more

  - Total number and kind of labels
    - **Large number of labels**
    - Specific labels

Human classification

➡ Already existing algorithm

➡ New algorithm developed externally

➡ New algorithm developed internally

A U T O M A T I C

Step 3: Choice of the classification method(s)

web data opp

1
2
3
4
5
6
7
8

- **Features of the tasks**

  - Total number of images
    - If low (<500)
    - If more

  - Total number and kind of labels
    - Large number of labels
    - **Specific labels**

➡ Human classification

Already existing algorithm

➡ New algorithm developed externally

➡ New algorithm developed internally

AUTOMATIC

Step 3: Choice of the classification method(s)

1
2
3
4
5
6
7
8

- **Features of the tasks**

  - Total number of images
    - If low (<500)
    - If more

  - Total number and kind of labels
    - Large number of labels
    - Specific labels

  - Recurrence of the study
    - **Punctual, or beginning of a recurrent collection**
    - Recurrent study

Human classification

Already existing algorithm

New algorithm developed externally

New algorithm developed internally

A U T O M A T I C

Step 3: Choice of the classification method(s)

web data opp

- **Features of the tasks**

  - Total number of images
    - If low (<500)
    - If more

  - Total number and kind of labels
    - Large number of labels
    - Specific labels

  - Recurrence of the study
    - Punctual, or beginning of a recurrent collection
    - **Recurrent study**

Human classification

Already existing algorithm

New algorithm developed externally

New algorithm developed internally

A U T O M A T I C

1

2

3

4

5

6

7

8

- **Resources**

  - Human resources
    - Manual classification: training of classifiers.
    - External algorithm: one person to check labels and one with technical knowledge.
    - Internal algorithm: highly specialized profile.

Human classification

Already existing algorithm

New algorithm developed externally

New algorithm developed internally

A U T O M A T I C

Step 3: Choice of the classification method(s)

1

2

3

4

5

6

7

8

- **Resources**

  - Human resources
    - Manual classification: training of classifiers.
    - External algorithm: one person to check labels and one with technical knowledge.
    - Internal algorithm: highly specialized profile.

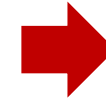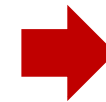 - Infrastructure
    - **In absence of proper hardware.**

➡ Human classification

➡ Already existing algorithm

New algorithm developed externally

New algorithm developed internally

A U T O M A T I C

Step 3: Choice of the classification method(s)

- **Resources**

  - Human resources
    - Manual classification: training of classifiers.
    - External algorithm: one person to check labels and one with technical knowledge.
    - Internal algorithm: highly specialized profile.

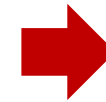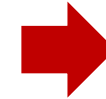 - Infrastructure
    - In absence of proper hardware.

  - Costs in terms of…
    - **Budget: if low…**
    - **Time: if limited…**

Human classification

➡ Already existing algorithm

New algorithm developed externally

New algorithm developed internally

AUTOMATIC

Step 3: Choice of the classification method(s)

web
data
opp

1

2

3

4

5

6

7

8

- **Resources**

  - Human resources
    - Manual classification: training of classifiers.
    - External algorithm: one person to check labels and one with technical knowledge.
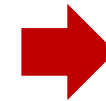   - Internal algorithm: highly specialized profile.

  - Infrastructure
    - In absence of proper hardware.

  - Costs in terms of…
    - Budget: if low…
    - Time: if limited…

  - Availability of images
    - **If images to train a new model are not available**

Human classification

Already existing algorithm

New algorithm developed externally

New algorithm developed internally

A U T O M A T I C

Step 3: Choice of the classification method(s)

- **Overall data quality**

- **Accuracy**

Human classification

Already existing algorithm

New algorithm developed externally

New algorithm developed internally

A U T O M A T I C

Step 3: Choice of the classification method(s)

web data opp

1

2

3

4

5

6

7

8

- **Overall data quality**

- Accuracy

- **Consistency**
  - **When using only one method**

Human classification

➡ Already existing algorithm

➡ New algorithm developed externally

➡ New algorithm developed internally

A U T O M A T I C

Step 3: Choice of the classification method(s)

web
data
opp

① ② ❸ ④ ⑤ ⑥ ⑦ ⑧

- **Overall data quality**

  - Accuracy

  - Consistency
    – When using only one method

- **Data protection**

➡ Human classification

Already existing algorithm

➡ New algorithm developed externally

➡ New algorithm developed internally

A U T O M A T I C

Step 3: Choice of the classification method(s)

- **Overall data quality**

  - Accuracy

  - Consistency
    - When using only one method

  - Data protection

- **Transparency**

Human classification

Already existing algorithm

New algorithm developed externally

New algorithm developed internally

A U T O M A T I C

1
2
3
4
5
6
7
8

Step 4: Collection of the images

- **Main considerations**

    - Getting respondents' consent to collect, analyze, archive and potentially share the images (when applicable)

    - Presenting images collected to respondents to check if they are good enough for the classification.

    - Processing sensitive data and metadata.

    - Size of the files during the collection.

    - Storage of images (safety criteria and files' size).

    - ID to relate to each participant

    - Tools to collect the images are necessary

1
2
3
4
5
6
7
8

web
data
opp

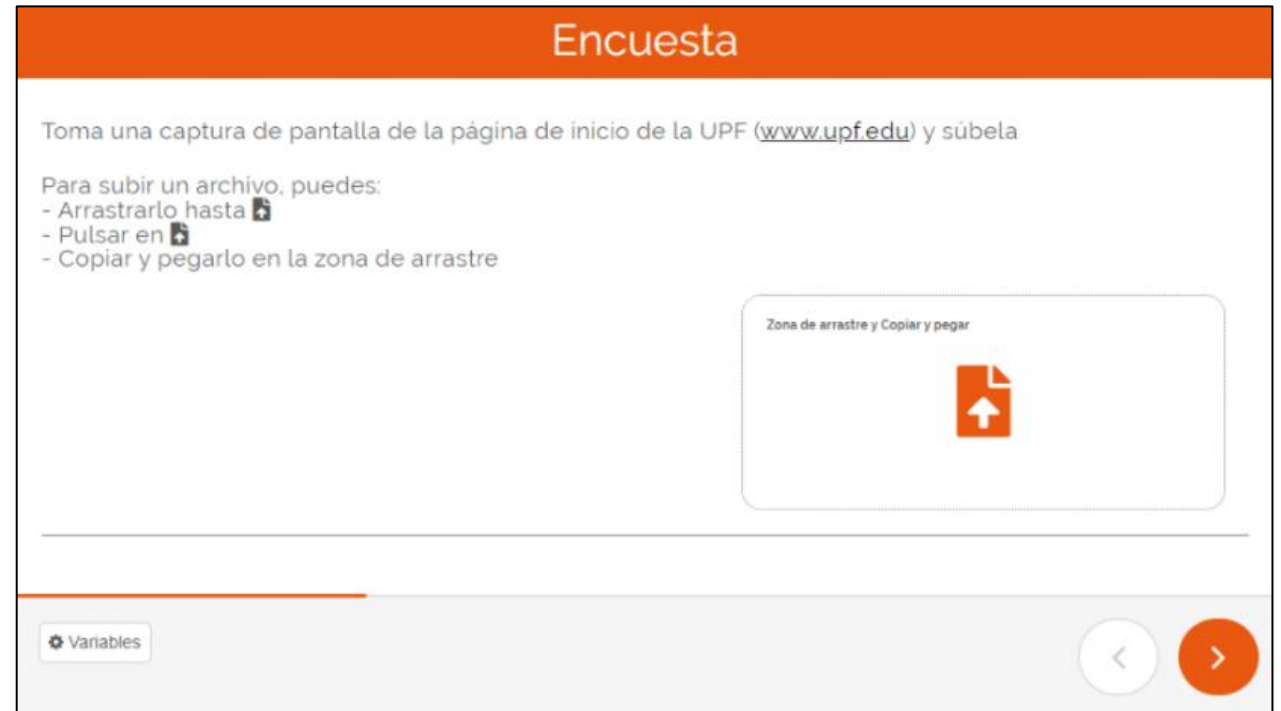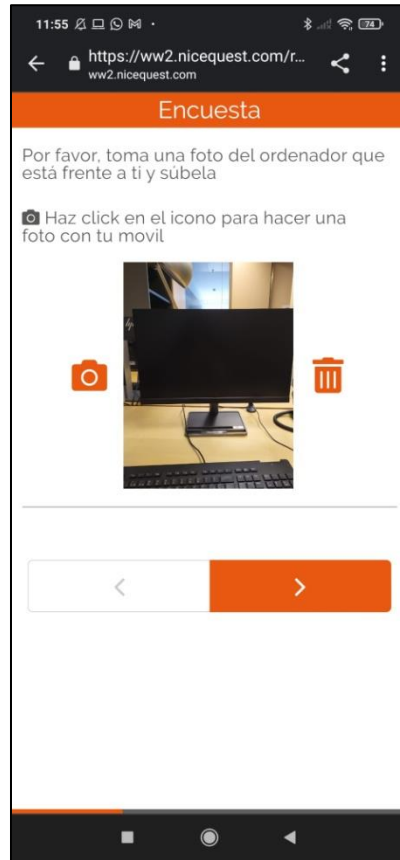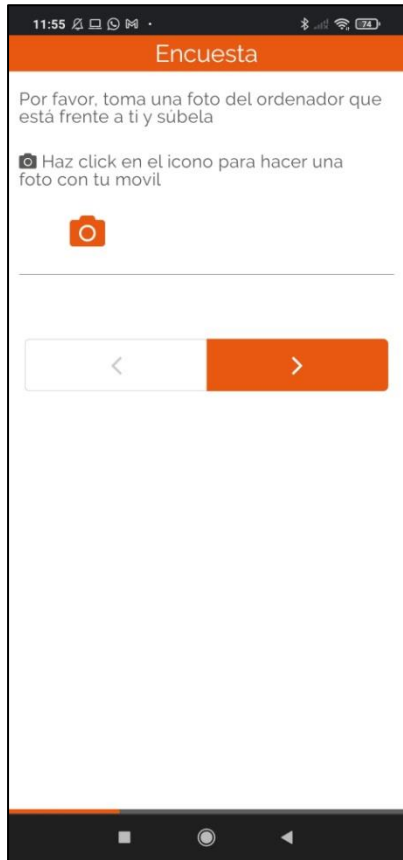# Step 4: Collection of the images

web data opp

- **Already available tools:** WebdataVisual (Revilla et al., 2022)

More information available at https://www.upf.edu/web/webdataopp/tools

Step 5: Images enhancement and revision

1

2

3

4

5

6

7

8

- **Images enhancement** → improvements in terms of lightning, sharpness, cropping…
  - When enhancement is not enough: discarding or re-collecting images

- **Ethical considerations** → blinding conflictive elements.
  - Revision by the ethical advisor and/or fieldwork company

Step 6: Classification

web data opp

- There are different tasks depending on the method(s) chosen:

1		
2	**Human classification**	• Classification guidelines • Training of classifiers
3		
4	Already existing algorithm	
5		
6	New algorithm developed externally	
7		
8	New algorithm developed internally	

Step 6: Classification

1

- There are different tasks depending on the method(s) chosen:

2

3

Human classification

- Classification guidelines
- Training of classifiers

4

Already existing algorithm

- Check the labels provided, and test the algorithm
- Upload images to the API or run the code
- Agree specificities of the dataset

5

6

New algorithm developed externally

7

8

New algorithm developed internally

Step 6: Classification

1

- There are different tasks depending on the method(s) chosen:

2

Human classification

- Classification guidelines
- Training of classifiers

3

4

Already existing algorithm

- Check the labels provided, and test the algorithm
- Upload images to the API or run the code
- Agree specificities of the dataset

5

6

New algorithm developed externally

- Definition of the requirements and performance goals
- Development of guidelines for the algorithm's parameters
- Potentially: provide images for training

7

8

New algorithm developed internally

web
data
opp

Step 6: Classification

1

2

3

4

5

6

7

8

- There are different tasks depending on the method(s) chosen:

Human classification	• Classification guidelines • Training of classifiers
Already existing algorithm	• Check the labels provided, and test the algorithm • Upload images to the API or run the code • Agree specificities of the dataset
New algorithm developed externally	• Definition of the requirements and performance goals • Development of guidelines for the algorithm's parameters • Potentially: provide images for training
New algorithm developed internally	• Have (or train) specialized programmers • Guarantee physical infrastructure • Provide images for training

Step 7: Verification of the classification outcomes

1

2

3

4

5

6

7

8

- **Within each method**

  - Human classification: swapping images between classifiers.

  - Automatic models: manual verification.

    Suggestions: swap at least 30 images to check accuracy, an acceptable threshold starts at 80% of accuracy (Zhang et al., 2022)

- **If using more than one method**

  - Contrast results between them.

- **Checking with participants**

  - Participants can access the labels and see if they are correctly classified.

  - Example: "cleansing gel" in a receipt.

Step 8: Analyses

1

- Researchers can perform different kinds of analyses, such as:
  - univariate analyses,
  - looking at the frequencies of each label,
  - the number of labels per image,
  - and other indicators that might help answering their research questions.

2

3

4

- The classified information can be **crossed with information from any other variables included in the survey** or that are provided by the fieldwork company (e.g., participants' profiling information).

5

6

- Overall, the analyses that can be performed with the labels are **very similar to those made with the codes of open-ended questions.**

7

8

Summary and conclusions

- When researchers are to collect images through an online survey, they can follow the eight steps proposed.

- **The process might not be linear**: it might require going back and forth across these steps.

- There is a **non-negligible risk of things going wrong:** respondents unable to send images or misunderstanding the request, losing the images due to storage problems, or obtaining biased results when pre-trained algorithms with inaccurate datasets.

- Thus, researchers should **carefully weigh the pros and cons before deciding to collect images.**

- Overall, working with images collected through online surveys is a demanding task, but it might be worthwhile when **images provide new and/or better insights, and improve the respondents' experience.**

Thanks!

PATRICIA A. IGLESIAS| Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra

patricia.iglesias@upf.edu

https://www.upf.edu/web/webdataopp

REFERENCES

Bosch, O., Revilla, M., & Paura, E. (2019). Answering mobile surveys with images: An exploration using a computer vision API. *Social Science Computer Review*, *37*(5), 669-683. https://doi.org/10.1177/0894439318791515

Ohme, J., Araujo, T., de Vreese, C.H., & Piotrowski, J.T. (2020). Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function. *Mobile Media & Communication*, *9*(2), 293-313. https://doi.org/10.1177/2050157920959106

Pataki, B.A., Garriga, J., Eritja, R., Palmer, J.R.B., Bartumeus, F., & Csabai, I. (2021). Deep learning identification for citizen science surveillance of tiger mosquitoes. *Scientific Reports*, *11*(1), 4718. https://doi.org/10.1038/s41598-021-83657-4

Revilla, M. (2022). How to enhance web survey data using metered, geolocation, visual and voice data? *Survey Research Methods*, *16*(1), 1-12. https://doi.org/10.18148/srm/2022.v16i1.8013

Revilla, M., Iglesias, P.A., Ochoa, C., & Antón, D. (2022). *WebdataVisual: A tool to collect visual data within the frame of web surveys* [Computer software]. OSF. https://doi.org/10.17605/OSF.IO/R7CAX

Zhang, R., Gong, J., Ma, S., Firdaus, A., & Xu, J. (2022). Automatic Coding Mechanisms for Open-Ended Questions in Journalism Surveys: An Application Guide. *Digital Journalism*, 11(2), 321-342. https://doi.org/10.1080/21670811.2022.2037006