Typology

A comprehensive, sharply-defined categorization system

* Comprehensive: Any object in the space will fit one (or more) types
* Well-defined types: Each type has a definition that allows systematic evaluation if whether an object is of that type
* Objective-ish: definitions should strive to be as objective as possible.. objective often cannot be achieved, but there are strategies to make subjective categories more objective.

e.g

- BAD: cookbooks, sci-fi novels..

- GOOD: fiction, non-fiction

Difficulties:

Hard to define (hate speech, pornography…)

Hard to discern Hard to determine from available information (Home location, Marital status)

Typology Design

A document consisting of…

* Motivation & context – why this typology needs to exist
* Overview of the types and their relation to one another
* List of types. For each type:
  + Concise definition
  + Positive examples with inclusion rationale
  + Negative examples with exclusion rationale
  + Edge cases with inclusion/exclusion rationale
* Argument or evidence for comprehensiveness

**Steps**

1. Get representative data

2. Get typology (find existing if it exists, or build your own using open coding)

3. Sanity check: evaluate typologies on representative data … can YOU make

them work?

4. Human test: Evaluate typologies on representative data with “expert” coders…

people you trust and believe can apply the typology as defined.

5. Does typology work? If yes, done! If no, adjust the typology and go to step 3

**Developing a typology through open coding**

1. Take a sample of data

2. Go through the sample and come up with categories

3. Review categories – are there any that are…

• Related or overlapping? Should these be merged?

• How “solid” is each? Assess whether it’s a real thing… could these fit in another category? Do we need this level of resolution?

• Are there any gaps (kinds of things that could happen, but you haven’t seen?) … go find some examples of these if you can.

**Comprehensive**

* Gather and look at extensive sample – if typology applies everywhere, chances are good it is near comprehensive.
* Catch-all category “other” – worst option

**Well-defined**

* Each definition should have rules for when they apply (and don’t)
* Each definition should be discernable from the data
* Make sure there is a (not too broad) way of handling ambiguous data.

**Objective-ish**

* Some types may be inherently subjective (beauty, goodness, acceptableness, etc…)
* Truly subjective categories are rarely useful – they will vary based on who you ask!
* To avoid subjective types, *ground the definition in a point-of-view*

• E.g., verifiability (assume answer as a journalist)

• E.g., edible (assume cultural context)

Confirm annotation quality

**Cohen’s kappa**

A way of scoring how good the agreement between two coders’ annotations are.

A close up of a clock

Description automatically generated

p0: frequency of agreement

p1: frequency of agreement by chance

Human coding

Ethical approval

* Harm to coders due to exposure
* Harm to others due to privacy

**Training**

Help coders understand the concepts and how they apply at a deep level.

• Prepare a code book (an extension to the typology definition

document)

• Prepare challenging examples

• Run group sessions with ample time to code content and discuss codings

**The human coding process**

Preparation phase

• Decide on human coding group (experts, small team, crowd)

• Develop code book/training materials

• Develop annotation software (if needed)

Annotation phase

• Recruit coders

• Run training session

• Run coding exercise (often spans days, sometimes weeks)

Validation phase

• Throughout, assess the quality and level of agreement among coders

Automated annotation

Diagram

Description automatically generated