

Extractor Instructions

2025-12-11 at 09:00:45 CET (UTC+0100)

Contents

Extractor instructions

Welcome! Welcome to the extraction instructions for EMPATHS-1, the first systematic review in the EMPATHS project. If this is new to you, you may want to start at <https://archeologists.openscience/empaths.html>. This PDF with extraction instructions is available from <https://archeologists.openscience/empaths-1/extractor-instructions>. In this project, the focus is on construct definitions and measurement methods. Therefore, during extraction, these are the main entities that you will spend time on. In addition to the brief extraction instructions specified in these instructions and in the extraction script (.Rxs file), where you will register the extracted data, more extensive instructions are provided here.

Please start by reading these instructions carefully, as well as the extraction instructions at <https://archeologists.openscience/empaths-1/extraction>. The instructions have two parts: this first part contains general instructions. The second part, starting from “Entity overview (list)”, contains entity-specific extraction instructions that will also be included in the Rxs Template (the “R Extraction Script Template”). That template is where you will conduct the extraction.

You will start any extraction by copying that template file to a new filename, and then opening that new file to enter the extracted information.

Naming the Rxs file The filename should follow this format:

“name_year_sourceId_extractorId.rxs.rmd”

Where ‘name’ is the first word in the last (family) name of the first author, stripped of all characters other than a-z or A-Z; ‘year’ is the year of publication of the source; ‘sourceId’ is the source’s unique identifier (the ShortDOI if available; otherwise, the QURID); and ‘extractorId’ is the extractor’s unique identifier (i.e., *your* identifier).

General extraction instructions To extract information from the sources, you scroll through the Rxs Template (that you just stored under a new name) and specify what you found for each entity.

You usually enter information by replacing the NULL with the entity content.

Usually, if something is not reported, replace the NULL with NA (also without quotes).

If you extract a number, you can usually just replace the NULL with that number. If you extract text, make sure to use double quotes around the text string.

Sometimes, you can extract multiple values (you can see this in the entity extraction instructions or in the instructions for the value template). In that case, you place them within a “concatenator” or “combiner”: c(). For example, c(1, 2, 3) for numbers, or c(“one”, “two”, “three”) for text strings.

Some clustering entities (i.e. containers of several entities that are closely clustered together) are repeating entities. This means you can copy them multiple times if need be. The empathy definition is an example:

a source can report multiple empathy constructs, and so require multiple versions of the empathy construct clustering entity for accurate extraction. Repeating entities can be recognized by the row of tildes (~) marking their beginning and end, as well as by the text “(REPEATING)” in their first and last lines. You can copy such a block, including the lines with the tildes, and paste the block immediately following the last line with tildes, ideally with one or more empty lines in between to clearly visually distinguish the blocks. You can repeat this process as often as necessary.

Validating your extraction results If you completed an R Extraction Script (.Rxs.Rmd file), you can (and should) immediately verify whether everything went well. There are two ways to do this. First, there’s the Extraction Validation App (EVA). Eva lives is at <https://opens.science/apps/eva>, and can validate your completed extraction script regardless of where you performed the extraction. Second, if you performed the extraction in RStudio, you can render the extraction script with CTRL-ALT-K. This will also produce the validation report, showing which entities validated and what is imported if your extraction script is parsed.

Extract construct definitions You extract the empathy definition in a so-called “clustering entity” or “list entity”: a set of closely related entities placed closely together in the Rxs template. If you don’t use RStudio for extraction (and so do not benefit of syntax coloring), this can look a bit confusing; clustering entities contain the entity itself (often with default value NULL), immediately followed on the same line by a comment (starting with three hashes, “###”) with the entity’s description, extraction instruction, and the corresponding value template description. Because this is a lot of text, editors (such as Notepad++ and RStudio) will often apply soft word wrapping (splitting long lines and displaying them over multiple lines to prevent them from disappearing off the right side of the screen). You may want to study this clustering entity closely the first time.

The first entity in this clustering entity is the Empathy Construct Identifier for this specific definition. This is used to allow multiple empathy constructs to be defined in (and extracted from) the same source. However, you have to specify an Empathy Construct Identifier even if a source only has one definition: in other words, you *always* have to specify it. Remember: identifiers can only contain letters, digits, and underscores, and must start with a letter.

The second entity in this clustering entity is the empathy definition. When looking for a definition of empathy, start by using your source viewer (e.g. if the source is in PDF format, it may be your browser (e.g. Firefox) or a dedicated PDF viewer such as Sumatra or Adobe Acrobat) and use the search/find functionality to look for the text string “empathy” (assuming the source was written in English). Ignore definitions in the abstract. If the first occurrence of the construct name is accompanied by its definition, as the authors use it in their work, copy that definition into the extraction script. An example of a (very very brief) definition you might encounter is “Empathy is the ability to understand and relate to the emotions and experiences of others and to effectively communicate that understanding.” (note that definitions can also be much longer).

However, if the first occurrence is accompanied by a definition that the authors discuss, but not as a definition they use themselves but rather e.g. to introduce readers to the definitions that exist, move to the next occurrence. Similarly, if the first occurrence of the word is not accompanied by a definition at all, move to the next occurrence. For each occurrence, repeat this evaluation: are the authors defining what exactly empathy is? In other words, which parts of the human psyche they consider constituting empathy, and which they consider to reflect other constructs?

Once you extracted the first fragment (i.e. one or more sentences), repeat your search to see whether the authors provide additional aspects of their definition further on in the introduction. If they do, extract those as well. Extract fragments that occur at different places on the text as separate text elements (e.g. `c("first bit", "second bit")`).

If the authors do not provide an explicit definition, then they may instead cite another source (e.g. an article or a book) and refer to the definition there as the one they use. In that case, obtain the shortdoi for that source, and extract that, in the full URL form (e.g. “<https://doi.org/gf6btx>”). This will enable us to later automatically identify all such URLs, and so categorize sources as either providing their own definition,

providing no definition, or citing a definition from elsewhere in the literature (as well as compile a list of such references). If they cite a source that does not have a DOI, consult with the EMPATHS-1 coordinators, Jennifer Gutsell and/or Gjalt-Jorn Peters.

If the authors do not define empathy but also do not cite another source as providing the definition they use, extract NA to signify that the definition is missing from the source. Similarly, if authors are not explicit about their definition, extract NA. If authors only provide a definition of empathy in the abstract, report that in the comments field in this clustering entity.

If a source is written in a language that you do not understand, extract “lang” as construct definition. This will allow us to later try to find somebody who *can* read that language.

Finally, some sources may contain multiple empathy constructs. In that case, extract them into separate entities. To do this, copy the block starting with the line containing “START: empathyConstruct (REPEATING)” and ending with the line containing “END: empathyConstruct (REPEATING)”. Then complete both entities for the second empathy constructs, and repeat until you extracted all different empathy constructs in the source. (An example of such a paper is ns9s; see <https://doi.org/ns9s> for the PDF and [URL] for the completed Rxs file.)

Extracting a measurement or manipulation instrument When extracting a measurement or manipulation (entities `empathyMeasureId` and `empathyManipulationId`), you specify their unique identifier. This identifier is taken from <https://archeologists.opens.science/empathy-measures> (from the column marked “identifier”). If the instrument you’re extracting is already in the list, you can just specify the relevant identifier in the extraction script.

However, if it does not yet exist, you have to add it. To do this, visit <https://opens.science/apps/elsa>, create an identifier, and add into the first column. Then specify the rest of the information as described in section “Specifying measurement instruments and/or manipulations” in <https://archeologists.opens.science/extraction>.

Just like definitions, a study can contain multiple measurement instruments or manipulation instruments. Again, copy the relevant block, from the line with “START: empathyMeasure (REPEATING)” to the line with “END: empathyMeasure (REPEATING)” for multiple measures, and from the line starting with “START: empathyManipulation (REPEATING)” to the line ending with “END: empathyManipulation (REPEATING)” for multiple manipulations.

Conversely, a study may not contain any measurement instruments or manipulations. In that case, you can specify “noMeasure” as value of `empathyMeasureId` and leave NULL as `empathyMeasureConstructId`, or “noManipulation” as `empathyManipulationId` and leave NULL as `empathyManipulationConstructId`.

Extracting multiple studies Sometimes, a source reports on multiple studies. If the studies use different measurement instruments or manipulation instruments, copy the study block like you may have copied definition blocks, measurement instrument blocks, or manipulation instrument blocks before. However, study blocks are larger, and themselves contain ‘repeating’ container entities (specifically, the measurement instrument blocks and the manipulation instrument blocks are specified within the respective study).

To copy the study block, copy the lines in between the line with “START: singleStudyContainer (REPEATING)” to the line with “END: singleStudyContainer (REPEATING)”. As you’ll see, this is quite a large part of the Rxs file. Also note that you may have to specify the population for each study separately.

How to create an identifier To create a unique identifier for a TOM, TOQ, or TOI, you can either use the R package {psyverse} or the Elsa app. To use Elsa, visit <https://opens.science/apps/elsa>. Identifiers follow the following format. They start with a brief lowercase sequence of letters that is often an acronym or abbreviation of the instrument’s name (e.g. ‘iri’, ‘bespt’, and ‘epitome’). This is followed by a number: the number of items in the measurement instrument; 0 for a manipulation; or 00 for continuous measurement such as EEG. That is followed by the language of the measurement instrument in ISO 639-3 code (see the

extraction instructions for extracting the language a source was written in). That is followed by an underscore, and then the last identifier bit as produced by Elsa.

Future reference In the future, we will specify the extracted measurement instruments and manipulations in an open repository. For that stage a number of instructions are included here. **During the extraction phase of this project, you can ignore this; this is simply retained here for future reference.**

If it is a questionnaire, you can choose to specify it as a TOQ (“Tabulated Open Questionnaire”) specification, enabling importing it into the questionnaire repository at <https://operationalizations.com>. This is not yet possible for measurement instruments that do not consist of questions and for manipulations; those have to be specified as TOM (“Tabulated Open Metadata”) specifications. Depending on what you choose, follow the corresponding set of instructions below.

Minimal specification of a measurement or manipulation instrument To specify a TOM (“Tabulated Open Metadata”) specification, you need to complete these steps:

- 1) visit <https://archeologists.opens.science/empathy-tabulated-specs>
- 2) open “TOM-spec—bespt0eng_7rtpjgf3”
- 3) save a copy under a different name but in the same folder.
- 4) create an identifier prefix (see the procedure below for details) and enter it in cell B3
- 5) visit <https://opens.science/apps/elsa>, enter the prefix, and create an identifier
- 6) enter the result in cell B4 as UMID
- 7) complete the other fields
- 8) open the spreadsheet at <https://archeologists.opens.science/empathy-measures> again and add a row with the UMID you just created

Full specification of a questionnaire To specify a TOQ (“Tabulated Open Questionnaire”) specification, you need to complete these steps:

- 1) visit <https://archeologists.opens.science/empathy-tabulated-specs>
- 2) open “TOQ-spec—eq60eng_7rs8g3bd”
- 3) save a copy under a different name but in the same folder.
- 4) create an identifier prefix (see the procedure below for details) and enter it in cell B3
- 5) visit <https://opens.science/apps/elsa>, enter the prefix, and create an identifier
- 6) enter the result in cell B4 as UQID
- 7) complete the other fields
- 8) open the spreadsheet at <https://archeologists.opens.science/empathy-measures> again and add a row with the UQID you just created

How to create an identifier To create a unique identifier for a TOM, TOQ, or TOI, you can either use the R package {psyverse} or the Elsa app. To use Elsa, visit <https://opens.science/apps/elsa>. Identifiers follow the following format. They start with a brief lowercase sequence of letters that is often an acronym or abbreviation of the instrument’s name (e.g. ‘iri’, ‘bespt’, and ‘epitome’). This is followed by a number: the number of items in the measurement instrument; 0 for a manipulation; or 00 for continuous measurement such as EEG. That is followed by the language of the measurement instrument in ISO 639-3 code (see the extraction instructions for extracting the language a source was written in). That is followed by an underscore, and then the last identifier bit as produced by Elsa.

Entity overview (list)

This is an overview of the entities to extract, their titles and descriptions, and other details that will become part of the extraction script template that will be used for the actual extraction.

General General information

Type: Entity Container

Identifier: general

Path in extraction script tree: source > general

Repeating: FALSE

QURID Quasi Unique Record Identifier (QURID).

Extraction instructions: This is already available in the tracking sheet; a QURID was added to every record. We will use this to automatically import bibliographic information available in that file, such as title, keywords, potentially abstract, etc.

Type: Extractable Entity

Identifier: qurid

Value description: A single character value that is used as an identifier and so is always mandatory and can only contain a-z, A-Z, 0-9, and underscores, and must start with a letter.

Path in extraction script tree: source > general > qurid

Value template: string_identifier

Repeating: FALSE

Language The language in which the article is written as ISO 639-3 code (e.g., to list the 10 most spoken languages: “eng” for English, “zho” for Chinese, “hin” for Hindi, “spa” for Spanish, “fra” for French, and “ara” for Arabic, “ben” for Bengali, “por” for Portuguese, “rus” for Russian, and “urd” for Urdu).

Extraction instructions: Use ISO 639-3 to extract this (see https://en.wikipedia.org/wiki/List_of_ISO_639_language_codes and https://en.wikipedia.org/wiki/ISO_639-3).

Type: Extractable Entity

Identifier: language

Value description: A single character value

Path in extraction script tree: source > general > language

Value template: string

Repeating: FALSE

Empirical Whether this source reports on one or more empirical studies (i.e. studies where data were created in the context of the study that authors report on, for example through experimentation, observation, simulation, or similar means).

Extraction instructions: Extract “yes” if this source reports results from at least one empirical study. Extract “no” if it does not report results from an empirical study. Extract “unclear” if you are not sure whether results from an empirical study are reported.

Note that since the species under study can be synthetic, collecting data from generative AI or a simulation also counts as empirical.

Type: Extractable Entity

Identifier: empirical

Value description: A string that has to exactly match one of the values specified in the “values” column of the Coding sheet, and that can be omitted (i.e. is allowed to be NULL).

Path in extraction script tree: source > general > empirical

Value template: categorical_omittable

Repeating: FALSE

Empathy Constructs This container entity is used to extract information about the various empathy constructs studied in this source.

Type: Entity Container

Identifier: empathyConstructs

Path in extraction script tree: source > empathyConstructs

Repeating: FALSE

Empathy Construct This clustering entity contains information about one single empathy construct as defined in this source. Note that we take a broad view of empathy constructs; this also includes empathy not as a part of the human psyche, but as it may be perceived to be expressed in, for example, a text or recording.

Type: Extractable Entity List

Identifier: empathyConstruct

Empathy Construct Identifier

This is a unique identifier for this empathy construct. It can be used elsewhere in this extraction script to refer to this construct (for example when extracting measurement instruments or manipulations).

Empathy Definition

The definition of empathy the authors use.

Empathy Definition Confidence

How confident you are that the definition you extracted is indeed how the authors defined empathy in this source.

Empathy Construct Type

The type of empathy construct: psychological construct or not.

Empathy Definition Notes

Any notes you want to specify.

Path in extraction script tree: source > empathyConstructs > empathyConstruct

Repeating: TRUE

Methods This container entity holds entities related to the methods used by the study.

Type: Entity Container

Identifier: methods

Path in extraction script tree: source > methods

Repeating: FALSE

Reported Studies This contained entity holds the studies reported on in this source.

Type: Entity Container

Identifier: reportedStudies

Path in extraction script tree: source > reportedStudies

Repeating: FALSE

Single Study This container entity contains information about a single study. This is important because some sources report on multiple studies.

Type: Entity Container

Identifier: singleStudyContainer

Path in extraction script tree: source > reportedStudies > singleStudyContainer

Repeating: TRUE

Population Information about the population of this study.

Type: Entity Container

Identifier: population

Path in extraction script tree: source > reportedStudies > singleStudyContainer > population

Repeating: FALSE

Species Whether the sample was drawn from humans or non-human populations

Extraction instructions: Extract “human” if the sample description in the methods section indicates a human sample. Extract “animal” if the description in the methods section of none of the studies reported indicates a human sample. Extract “synthetic” if the data were produced by an automated algorithm (e.g. a simulation such as a large language model or an agent-based model). If another species was studied, extract “other” and then also specify that species in the “population_species_other” entity. If the collected data was produced by multiple species, extract all species as a vector (see the examples).

Type: Extractable Entity

Identifier: population_species

Value description: A vector of strings where each element has to exactly match one of the values specified in the “values” column of the Coding sheet

Path in extraction script tree: source > reportedStudies > singleStudyContainer > population > population_species

Value template: categorical_multi

Repeating: FALSE

Other Species If the species that was specified was “other”, then as this entity extract the text fragment where the authors describe the species they studied.

Extraction instructions: Extract the literal text the authors use; if the species was not extracted as “other”, extract this as NA.

Type: Extractable Entity

Identifier: population_species_other

Value description: A single character value; can be NA or even NULL

Path in extraction script tree: source > reportedStudies > singleStudyContainer > population > population_species_other

Value template: string_omittable

Repeating: FALSE

Manipulation Whether the source involves a manipulation of empathy (or intervention, behavior change method, therapy component, etc).

Extraction instructions: Assess whether the source introduces or involves a procedure designed to increase, decrease, or otherwise alter the research units’ empathy (i.e. the humans or animals that are studied). This can be called a manipulation in experimental psychology, a behavior change method, technique or principle in behavior change science, or a therapy component in clinical psychology. Other terms are also possible of course: the key is whether the procedure or stimulus was designed to influence empathy. If you conclude that such a procedure or stimulus is described in the source as one of the focal topics, extract “yes”. If you conclude that no such procedure or stimulus is described, extract “no”. If it is unclear whether that is the case, extract “unclear”. If nothing is reported that allows you to draw any conclusions, extract NA (without quotes).

Type: Extractable Entity

Identifier: involvesManipulation

Value description: A string that has to exactly match one of the values specified in the “values” column of the Coding sheet, and that can be omitted (i.e. is allowed to be NULL).

Path in extraction script tree: source > reportedStudies > singleStudyContainer > involvesManipulation

Value template: categorical_omittable

Repeating: FALSE

Empathy Measures This container entity holds entities specifying how empathy was measured.

Type: Entity Container

Identifier: empathyMeasures

Path in extraction script tree: source > reportedStudies > singleStudyContainer > empathyMeasures

Repeating: FALSE

Empathy Measure Container entity for this empathy measure.

Type: Extractable Entity List

Identifier: empathyMeasure

Empathy Measure Identifier

The identifier for the empathy measure that was used to measure empathy in this study in this source.

Measured Construct

The identifier of the construct as entered in its extracted definition above.

Measure notes

Any notes on this measurement instrument.

Path in extraction script tree: source > reportedStudies > singleStudyContainer > empathyMeasures
> empathyMeasure

Repeating: TRUE

Empathy Manipulations This container entity holds entities specifying how empathy was manipulated.

Type: Entity Container

Identifier: empathyManipulations

Path in extraction script tree: source > reportedStudies > singleStudyContainer > empathyManipulations

Repeating: FALSE

Empathy Manipulation Container entity for this empathy manipulation.

Type: Extractable Entity List

Identifier: empathyManipulation

Empathy Manipulation Identifier

The identifier for the empathy manipulation that was used to manipulate empathy in this study in this source.

Manipulated Construct

The identifier of the construct as entered in its extracted definition above.

Manipulation notes

Any notes on this manipulation

Path in extraction script tree: source > reportedStudies > singleStudyContainer > empathyManipulations
> empathyManipulation

Repeating: TRUE
