Environmental Evidence

METHODOLOGY

Open Access

Novel tools and methods for designing and wrangling multifunctional, machine-readable evidence synthesis databases

Neal R. Haddaway[1,2,3*], Charles T. Gray[4] and Matthew Grainger[5]

Abstract

One of the most important steps in the process of conducting a systematic review or map is data extraction and the production of a database of coding, metadata and study data. There are many ways to structure these data, but to date, no guidelines or standards have been produced for the evidence synthesis community to support their production. Furthermore, there is little adoption of easily machine-readable, readily reusable and adaptable databases: these databases would be easier to translate into different formats by review authors, for example for tabulation, visualisation and analysis, and also by readers of the review/map. As a result, it is common for systematic review and map authors to produce bespoke, complex data structures that, although typically provided digitally, require considerable efforts to understand, verify and reuse. Here, we report on an analysis of systematic reviews and maps published by the Collaboration for Environmental Evidence, and discuss major issues that hamper machine readability and data reuse or verification. We highlight different justifications for the alternative data formats found: condensed databases; long databases; and wide databases. We describe these challenges in the context of data science principles that can support curation and publication of machine-readable, Open Data. We then go on to make recommendations to review and map authors on how to plan and structure their data, and we provide a suite of novel R-based functions to support efficient and reliable translation of databases between formats that are useful for presentation (condensed, human readable tables), filtering and visualisation (wide databases), and analysis (long databases). We hope that our recommendations for adoption of standard practices in database formatting, and the tools necessary to rapidly move between formats will provide a step-change in transparency and replicability of Open Data in evidence synthesis.

Keywords: Automation, Data abstraction, Evidence synthesis technology, Living systematic reviews, Systematic map, Systematic review

Background

Why databases are integral to evidence syntheses

One of the most important steps in the process of conducting a systematic review or map is data extraction: locating and abstracting information and study findings from within each manuscript and entering these data into a specifically designed database. Methodological guidance for systematic reviews and maps advocates that these databases should be predesigned and planned from the outset to maximise consistency, efficiency and objectivity in how data are extracted [1]. However, systematic review authors (referred to hereafter as evidence 'synthesists') typically design databases from scratch for each review project, and there has been no attempt to standardise systematic review data formats. Whilst systematic review management tools (e.g. CADIMA; [2]) each have their own standard format for storing and exporting data extracted within a systematic review or systematic map, there is no effort to ensure consistency across tools.

*Correspondence: neal_haddaway@hotmail.com; neal.haddaway@sei.org
[1] Stockholm Environment Institute, Linnégatan 87D, Stockholm, Sweden
Full list of author information is available at the end of the article

Furthermore, there is little adoption of easily machine-readable, readily reusable and adaptable databases. Providing digital versions of review or map does not mean that these data are easily understandable, or provided in a format that can be readily reused without considerable time and effort needed to 'wrangle' the data into a usable format. Review and map databases that are provided in a consistent and clear format that obeys certain standards and rules would be easier to translate into different formats: not only by review authors themselves (for example when producing narrative tables, visualisations and continuing to data analysis) but also by readers of the review/map if they wish to interrogate the data, replicate the analyses or reuse data for other purposes.

There are many ways to structure data, from graph databases [3], delimited text format, to json files [4]. As evidence synthesists, it is often not possible to select a database format that will be optimal for all use cases. Instead, we suggest that the structure the data takes should be carefully considered and planned. Our purpose here is to delineate between **structured** (i.e. readily machine readable) and **visualised** (i.e. human readable) data. We believe that both formats are important; and, in particular, stress that human readable tables are not sufficient for computational reproducibility and interoperability.

Efforts have been made to define best practice in computational work with data; a researcher may adhere to, for example, FAIR (findable, accessible, interoperable, and reusable) principles [5], or the TRUST (transparency, responsibility, user focus, sustainability, technology) principles for digital repositories [6], but precisely how to implement these principles with the tools available for a given discipline is left to the researcher.

As the drive towards Open Science and Open Synthesis picks up pace [7, 8], consistency in the way the Open Data (the immediate publication of full research data free-of-charge; [9]) are presented is important to maximise transparency, usability and legacy of data from evidence syntheses.

In evidence synthesis, there are very sensible reasons to structure a data table by study, so that a reader can easily read and summarise vertically, horizontally, between studies, and within sub-tables. This structure is optimised for the reader, but it is a significant barrier to the extensibility or reuse of the research, now considered a best practice in scientific computing [10]. Wilson et al. [11] suggest that we should aim for '*good enough*' practices. As such, it is likely a researcher reading this manuscript would not be trained in databases, but, like the authors, would be an *evidence synthesist*, wanting to ensure they practice science diligently. In order to bridge this toolchain (i.e. a set of programming tools) gap between best practices in scientific computing and evidence synthesists' practice, we provide some examples of data structures and tools. Our central focus is to underscore that the choice of data structure for publication should be a fundamental component of the review project planning—outlined in the review protocol—and authors should ensure that: (1) the data are provided in a machine-readable format with human-readable summaries; and (2) the structure is clearly documented (i.e. the data provided are described and explained in detailed meta-data).

Why we should care about good data curation and Open Data

The Open Science movement has been instrumental in raising awareness and interest in the benefits of Open Data across disciplines [12]. Open Science has recently been applied specifically to evidence synthesis via the concept of Open Synthesis [7]. Various benefits to Open Data have been cited, including: increasing opportunities for reuse and further analysis (reducing research waste; [13]); facilitating real-time sharing and use of data [14]; increasing research visibility, discovery, impact and recognition [15]; facilitating research validity through replication and verification [16]; decreasing the risk of research fraud through transparency [17]; use of real research in educational materials [17]; facilitating collaborative research and reducing redundancy and research waste across siloed groups [18]; enabling public understanding [18]; increased potential to impact policy [19]; promotion of citizen science [20]. These benefits are the same for Open Data in evidence synthesis (i.e. systematic reviews and maps). Indeed, synthesists are often plagued by incomplete reporting of data and meta-data in primary studies [21]: we should therefore be keenly aware of our mandate to ensure we comply with Open Data principles for the same reasons.

Systematic reviews and maps involve the management of large, complex datasets, often with multiple dependencies and nesting: for example, multiple outcomes for a single intervention, multiple interventions within a single study, multiple studies within a single manuscript, and multiple manuscripts on single studies. Added to this, synthesists must comply with strict demands for rigour in the way data extraction is planned, executed and reported (e.g. the MECCIR standards for conduct of systematic reviews; https://onlinelibrary.wiley.com/page/journal/18911803/homepage/author-guidelines). Careful data curation in evidence synthesis is vital to avoid errors that easily propagate and threaten the validity of these substantial projects: particularly where multiple synthesists are working on the same evidence base simultaneously and over long time periods that frequently involve

staff turnover. Where used, systematic review management tools (e.g. SysRev.com) have come a long way to reduce unnecessary errors, but there remains no consistent approach to data extraction across platforms that hampers Open Synthesis [7].

Objectives

As described above, most systematic reviews and maps design bespoke data extraction tools. These databases obviously share similarities, for example citation information, study year, location, etc., and there are necessary differences depending on the focus of the review. However, each database uses different conventions related to column names, cell content formatting and structure. Because of the lack of templates across the evidence synthesis community, synthesists often learn the hard way that databases designed for data extraction are often not suitable for immediate visualisation or analysis. In addition, data cannot be readily combined across databases, meaning that overlapping reviews cannot share resources in data extraction.

We outline below several common problems with systematic review and map databases, and then go on to describe how adopting a standard template for database design can help synthesists to wrangle their data automatically for rapid visualisation and analysis, whilst also facilitating Open Data principles. We have the following objectives that sit within a broader theory of change related to improved data management in evidence syntheses:

- to provide recommendations in data formatting and propose a methodology for designing evidence synthesis databases.
- stimulate the production of tools to allow rapid reformatting of data between types suitable for reading and types suitable for analysis.
- to support Open Synthesis as a community, facilitating synthesis and reuse of syntheses.
- improve the legacy and impact of syntheses in the future.

Here, we provide one solution to the current problem of messy data: '*structured*' databases, in which data are shared in tables wherein each row denotes an observation and each column a variable [33], and a series of context-agnostic tools to *translate* databases between different formats for specific uses. Although we focus here on systematic review and map databases, our tools and recommendations are equally applicable to any form of evidence synthesis and usable in any form of database.

There are other options, for example, relational databases and graph databases [22]. Many of these solutions will work well together (e.g. relational databases can be based on these principles), but what we provide here is an easily understandable concept that does not require specialist software or knowledge of databases. It is also based on principles of interoperability with the most common statistical and programming language, R [23]. By building a database using the principles described herein, users can effortlessly move from a data extraction phase into a data synthesis phase (particularly for quantitative synthesis) without the need for manual (time consuming) data manipulation. This solution also allows a single database to be used for multiple types of synthesis; e.g. visualisation in flow diagrams, heat maps, evidence atlases and diagnostic plots as well as data analysis in meta-analysis. Although our examples are provided in the R coding environment and therefore require some basic knowledge of coding, we also intend to provide user interfaces that allow translation of databases with no prior coding experience necessary.

We begin by introducing common formats for evidence synthesis databases, and then describe how databases have been designed and published in Environmental Evidence. We then outline how existing data science tools can support efficient and transparent database design, translation and use. Finally, we provide recommendations and methods for designing databases and introduce tools for rapidly converting between formats for different purposes.

Evidence synthesis data formats

Broadly speaking, there are four formats for systematic review and map databases: '*condensed*' formats, '*wide*' formats, '*long*' formats, and '*wide-and-long*' formats (see Fig. 1).

Condensed formats are easy to comprehend, often with nested column headers to denote related variables. We have termed these 'condensed' because they are used in an attempt to both visualise the dataset in a tabular format for the reader and contain sufficient information to be a comprehensive dataset. They typically revolve around the principle of one-line-per-study, such that the study is implied to be the level of data independence. However, these databases are not readily machine readable (i.e. they cannot be immediately filtered, visualised or analysed) where multiple values exist in one cell, for example where studies contain multiple outcomes, and particularly where these outcomes were measured in different ways.

Where a database contains multiple columns each with multiple values per cell, the implicit linkages between cell values are assumed to apply for all values of all cells: in the example in Fig. 2, each of the three values in column

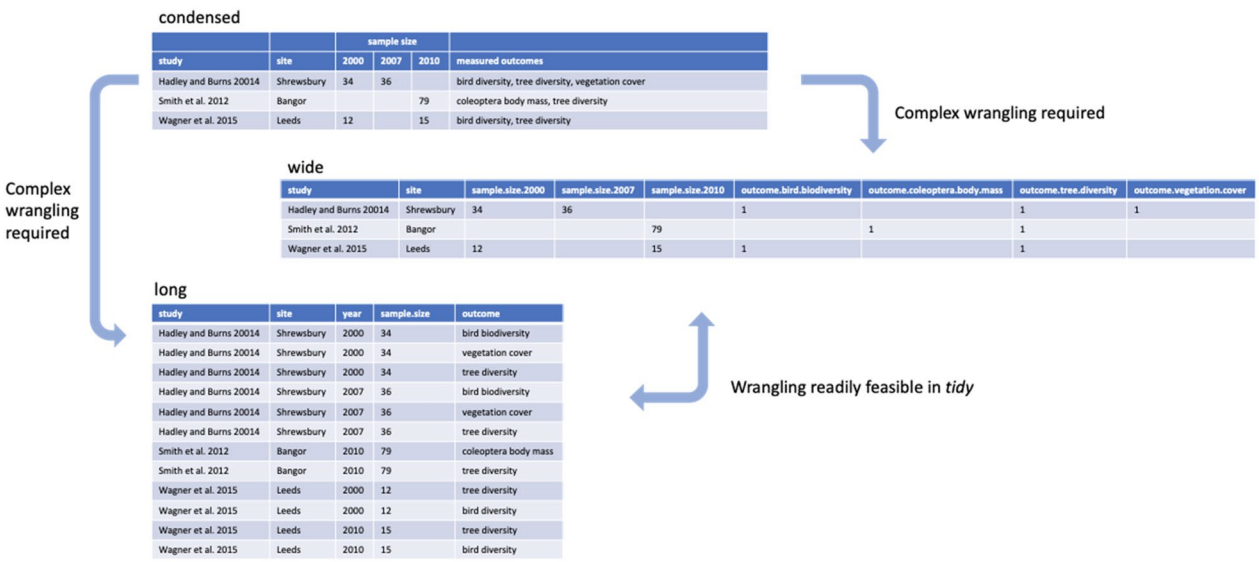

Fig. 1 Conceptual model of condensed, wide and long database formats in systematic reviews and maps

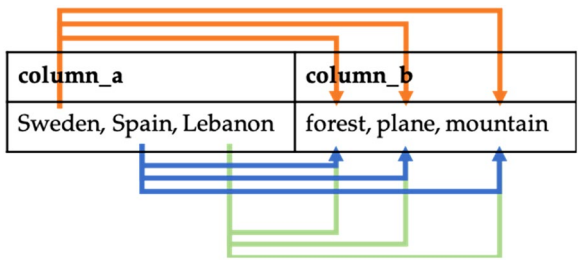

Fig. 2 Multiple values held within a single row across two columns make it impossible to discern implicit linkages between specific values. All values in one column are linked with all values in the other

A are associated with each of the three values in column B:

Condensed formats are perhaps the most common in CEE reviews (see "The status quo in environmental evidence syntheses" section for an analysis of recent reviews).

Wide format databases also typically consist of each row as a study, but avoid multiple values in a single cell that hamper filtering and linking between variables by separating different levels of a variable across multiple columns. Where a database has 10 variables extracted for each study, and each variable has 5 possible values, the database would require at least 50 columns for these data. In order to clearly denote which columns related to the same variable, some form of nesting is required: visually this could be done by having a hierarchy of column headers, but this cannot be readily read by a machine. Better yet, column names should follow standard naming conventions to display this nesting: for example, the variable 'colour' could be split across three columns named 'colour.yellow', 'colour.red', and 'colour.blue'. These naming conventions help to identify nesting by a human and machine reader. Despite not having multiple values per cell, wide formats can still suffer from linkage issues where multiple independent data are coded within a single line.

Long format databases obey the principle of each row being an independent data point. This is typical where a study has only one independent set of data (e.g. one method, one outcome, one time period), but will often involve multiple rows per study. Long databases need to track the relationships between rows carefully, and typically do this by using identification codes to denote a particular article, study, location, etc. within which multiple data points are nested. Careful naming conventions are still important for these databases, but each column will contain all levels of a variable, with multiple levels split across different lines, so naming is simpler.

Wide-and-long formats are a combination of wide and long database formats: for example, they will have a single line per outcome but different factors of a variable will be split across multiple columns.

The various advantages and disadvantages of each database format are outlined in Table 1.

Table 1 Advantages and disadvantages of the main database formats for evidence syntheses

data.format	Advantages	Disadvantages
Condensed	Easier to read and navigate (smaller, easier to scan, fits in a manuscript) Easy to enter data into Easy to understand (nested columns show relationships) No unnecessary blank spaces (avoids confusion)	Cannot be fed directly into visualisation or analysis Difficult to convert to wide or long format Easier to make/mask errors because of compressed information Cannot link multiple values across columns Cannot filter by multiple cell values
Wide	Avoids repeated information Easier to understand from a review perspective (one line per study) May feel easier to fill in (left-to-right) Easy to filter Easy to convert to long format	Requires careful naming conventions to demonstrate column nesting May be difficult to understand columns Lots of blank spaces Difficult to read on a landscape screen Requires wrangling for most visualisations and data analysis
Long	Easier to read than wide Designed for immediate visualisation and data analysis Easy to filter Easier to understand columns than wide format Easy to convert to wide format	Contains considerable repetition Difficult to read and navigate May be harder to fill in (requires a lot of scrolling) Does not require complex naming conventions (no nesting) No unnecessary blank spaces
Wide-and-long	Combines the advantages of both wide and long Enables between and within variable comparison Computationally efficient Closer to a conventional human readable table, as shown in Fig. 1	Embedded variables can be difficult to extract Embedded variables are frequently time consuming to extract

The status quo in environmental evidence syntheses

We undertook an analysis of all systematic reviews and maps published by the Collaboration for Environmental Evidence between May 2012 and May 2019 to investigate how synthesists have structured their databases and to what extent the information is reusable. We identified all reviews and maps by hand searching the journal Environmental Evidence (on 05/05/19). We then examined each review and extracted the meta-data outlined in Table 2. This information was checked by a second reviewer.

We identified a total of 29 systematic reviews and 18 systematic maps (see Additional file 1). A total of 44 of these 47 syntheses provided a digitised database of

Table 2 Meta-data extracted from CEE systematic reviews and maps published between May 2012 and May 2019

column_title	Description
review_title	Title of the review (with hyperlink)
Citation	journal_name year volume:article number
pubilcation_date	Date of publication according to EEJ
review_type	Systematic review or systematic map
database_provided	Was a database of studies and descriptive information provided as an additional file? 1 = yes, 0 = no
file_format	The format of the database file
data_as_rows_variables_as_columns	Is the data organised such that independent data points (e.g. studies) are rows and variables are columns? 1 = yes, 0 = no
multiple_values_per_cell	Are there any occurrences where cells contain more than one value? 1 = yes, 0 = no
potential_linking_issue	Are there any occurrences where cells contain more than one value across multiple variables? 1 = yes, 0 = no
format_condensed	Is the database arranged in a condensed format? See manuscript for full description? 1 = yes, 0 = no
format_wide	Is the database arranged in a wide format? See manuscript for full description? 1 = yes, 0 = no
format_long	Is the database arranged in a long format? See manuscript for full description? 1 = yes, 0 = no
format_wideandlong	Is the database arranged in a wide-and-long format? See manuscript for full description? 1 = yes, 0 = no
format_multiple	Are multiple database formats provided? 1 = yes, 0 = no
value_separator	Data separator used to indicate break between multiple values in a cell
multiple_separators_per_table	Are different separators used within the same data table? 1 = yes, 0 = no
multiple_separators_per_cell	Are different separators used within the same column? 1 = yes, 0 = no
Notes	Further comments on data formatting
database_link	The URL for the main database file

studies (either as a map database or as a table of data used in quantitative synthesis): only 2 databases did not employ a studies-as-rows principle, instead opting for studies as columns. Of the 44 synthesis databases, 32 were in a spreadsheet format, 5 were Microsoft Access databases, 3 were PDFs (portable document format), and 4 were Microsoft Word documents. A total of 28 syntheses used condensed format databases, 3 were wide, 14 were long, 2 was wide-and-long, 4 were relational formats, and 5 provided multiple formats (syntheses that used multiple formats are also counted in individual categories) (see Fig. 3).

We noted several other issues that would cause problems in machine readability, most commonly that 11 syntheses used multiple different separators for cell values, for example combining commas and semicolons in the same database. Furthermore, 6 databases included multiple separators in the same column. In one case, coding was used inconsistently within the same column, with the same codes being described using different grammar and spelling [24]. Some authors avoided multiple cell values by using 'multiple' as a cell content, with inherent loss of information (e.g. [25, 26]). One synthesis provided the database in a Microsoft Word format with its contents available as image files, and thus entirely non-machine readable [27].

Human readability is classified as both an advantage and disadvantage. Data structure for human interpretation has benefits for summary between and within studies, but it poses significant disadvantages for incorporating into future analyses and codeflows. Extracting embedded variables that have been summarised into one column or row presents an ongoing obstacle for evidence synthesists. The combined use of value separators makes it very challenging to extract single values from a column of a condensed database, particularly where responses are free text strings and may contain common value separators, like commas, semicolons or slashes.

A *structured* revolution for systematic review and map data

How to structure data in systematic reviews/maps

Systematic review authors can easily integrate standard structured data principles when designing and publishing their databases. We provide the following advice on designing and using databases within systematic reviews and maps:

- Systematic review authors should consider publishing databases of descriptive information about the included studies (similar to a systematic map database; a list of included studies along with their *metadata*—i.e. not just the study findings but all other extracted descriptive information alongside each study citation): this is rarely done but increases the utility of the review's contents, by allowing users to learn about the nature of the evidence base as well as its findings.
- Provide detailed explanatory notes that describe the column titles, and the database structure and conventions in a data dictionary file or tab of the database worksheet. Within reason, this information should ideally be sufficient to understand the data without needing to carefully read the manuscript.

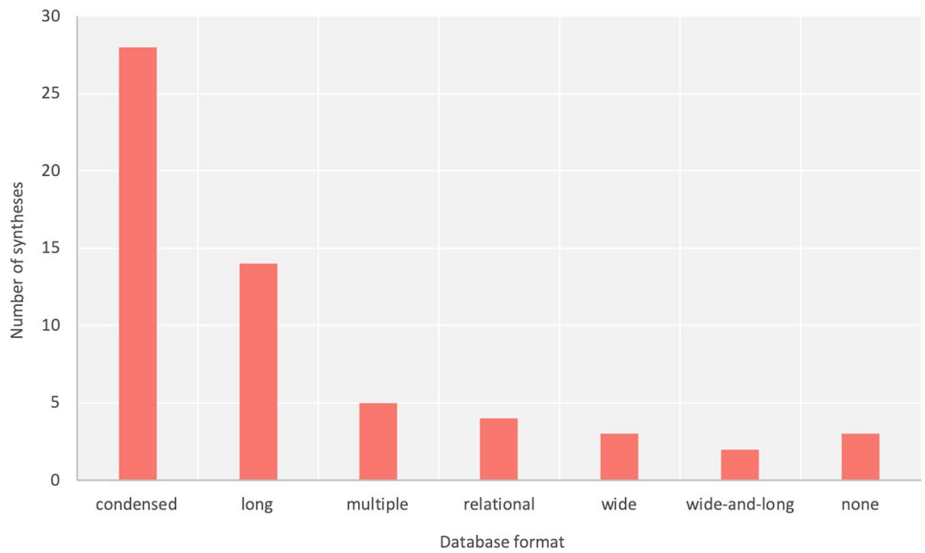

Fig. 3 Database formats used in systematic reviews and maps published in Environmental Evidence between 01/05/12 and 04/05/19. Syntheses that used multiple formats are also counted in individual categories

- Clearly explain what level of independent data has been used in the database: i.e. what is each row in your dataset (a manuscript, a study, a location, an outcome?).
- Facilitate computational transformation (i.e. immediate use of the database in a visualisation or statistical programme) by avoiding multiple values per cell and by using appropriate naming conventions for column names (see Box 1 and "snake_case naming convention for column/variable names", below).
- Where there is a need for linkages between multiple values across multiple cells, separate data onto different lines and describe this clearly.
- Become familiar with data wrangling tools to transform between wide and long formats.
- Consider providing multiple formats of a database that allow rapid reading by a human and immediate machine readability.

Translatable data

Standardised formats for database design and naming conventions

Whilst there are a range of naming conventions and options for database structures in systematic reviews, we propose the following practices to act as a standard across the community to maximise efficiency, reuse and training.

Differentiate between databases for visualisation and databases for analysis

The evidence synthesis community should delineate between making databases visually appealing for rapid understanding, versus databases that can be reused and immediately machine read. This should not require double the work: here, we advocate for a single, planned database designed for data extraction that can be wrangled into a variety of formats for different purposes using computational methods.

Synthesists should consider including both visual databases for human-readable tables and machine-readable data tables for analysis and replication: synthesists are already encouraged to provide condensed tables in a narrative synthesis (Collaboration for Environmental Evidence 2018). However, we believe that all tables should be provided in a digital, machine readable format as well as providing in-text tables to maximise legacy and impact of their work.

Sample conversions:

		Sample size			
Study reference	Study site	2000	2007	2010	Measured outcomes
Hadley and Burns 20014	Shrewsbury	34	36		bird diversity, tree diversity, vegetation cover
Smith et al. 2012	Bangor			79	coleoptera body mass, tree diversity
Wagner et al. 2015	Leeds	12		15	bird diversity, tree diversity

study	site	sample_size_2000	sample_size_2007	sample_size_2010	outcome_bird_biodiversity	outcome_coleoptera_body_mass	outcome_tree_diversity	outcome_vegetation_cover
Hadley and Burns 20014	Shrewsbury	34	36		1		1	1
Smith et al. 2012	Bangor			79		1	1	
Wagner et al. 2015	Leeds	12		15	1		1	

Principles of naming conventions to facilitate machine readability:

naming.convention	description	advantages	disadvantages
full.stop.separated	Spaces are replaced with full stops	The default conversion in R, relatively easy to read	In some coding languages, full stops are programmatically meaningful
underscore_separated	Spaces are replaced with underscores (aka snake_case)	Faster to read than other formats	Harder to type than full stops
lowerCamelBack	The first letter is lower case, other words are capitalised	Shorter than separated conventions, less prone to reading errors	Difficult to remember, harder to type than full stops
UpperCamelBack	All words are capitalised	Consistent, relatively easy to read, shorter than separated conventions, less prone to reading errors	Difficult to write quickly
alllowercase	All words are lower case	Shorter than separated conventions	Difficult to read

Example of a study comparing suitability of naming conventions/identifier styles:
Sharif, B. and Maletic, J.I., 2010, June. An eye tracking study on camelcase and under_score identifier styles. In *2010 IEEE 18th International Conference on Program Comprehension*(pp. 196-205). IEEE.

Box 1 How naming conventions can support tidy databases for evidence syntheses

Avoid multiple values per cell

Along with human-readable tables, synthesists should also provide either a wide or long (or wide-and-long) format database as an Additional file 1. There should ideally be a move away from the preference for providing only a condensed format that compresses multiple values into a single cell. Despite being visually appealing and easy to populate, these are difficult to interact with.

`snake_case` *naming convention for column/variable names*

'Naming conventions' are consistent ways of naming variables in a database that make it easier to recognise and use variable names in software languages like R and Python. In essence these are sets of rules for describing variables that cannot be easily named using single unique words or short strings. They are necessary because spaces within a variable name cause problems for tools that automatically detect where one entity stops and another begins. For example, the variable "study location" would ideally be converted into a single string. This could be done using a range of separators to make it easier for the human coder to use: studloc, study.location, study_location, studyLocation, study-location, etc. Naming conventions also facilitate automatic wrangling of data and make it easier to produce Open Source software to switch between database formats.

There are many different naming conventions; for example, `camelCase`, `snake_case`, `kebab-case`, `SCREAMINGCASE`, and combinations thereof [28]. Our examples below focus on the statistical programming language R, in which there is a convention of separating words with a full stop: for example, `read.csv`. However, this creates problems in other languages; in Python, full stops have additional language-specific functionality and create problems when used as variable names. Thus, current best practice is to use `snake_case`—lower case and separate words with an underscore [29]—thus improving interoperability.

- Consistency in whichever approach is chosen

  Whatever database format and naming convention is chosen, synthesists should ensure that they use this consistently and do not use multiple formats within a single database (e.g. combining naming conventions or being inconsistent in the use of wide and long formats).
- Develop and use tools to translate between database formats

We have developed a series of tools to support database translation in line with the principles described herein. We also provide a number of 'toolchain walkthroughs' combining existing functions together that demonstrate how they can be applied in practice to existing databases (https://softloud.github.io/sysrevdata/). We describe the relative roles of each format and example code for converting between them here. These methods combine existing code to wrangle data between formats commonly seen in systematic reviews and maps. The code is freely accessible and adaptable (i.e. published under CC-BY license), and may also be a useful basis for the production of Open Source tools with user interfaces to maximise usability by those with limited coding skills.

In our toolchain walkthroughs, we provide examples of restructuring data to illustrate the nuances and challenges of effectively sharing data for evidence synthesis. Communicating data effectively with collaborators and other scientists is not easy, see, for example, a recent retraction (https://laskowskilab.faculty.ucdavis.edu/2020/01/29/retractions), time for the development of required research software must increasingly be factored into the research plan. That being said, *best practice* in scientific computing is so challenging that best practices was updated to the more realistic *good enough* practices [11]. We are thus not claiming that our solutions are perfect, but rather fit-for-purpose, and act as a vital starting point for further development and optimisation. No doubt, code may need adapting for specific contexts, particularly where data has not been structured efficiently from the start.

Here, we provide a few examples (see the walkthrough for all examples) to give a sense of the type of problems one might encounter when structuring data for computational evidence synthesis. These examples are visualised for clarity in Fig. 4 with the example from Fig. 2.

Example 1 Separating multiple values per cell onto separate lines (condensed-to-long)

The following function makes use of the `separate_rows()` function in the 'tidyr' package [30] to split rows where a specific column has multiple values per cell into different lines. The default separator used below is '|||', as used in the free-to-use review management tool SysRev (http://sysrev.com). This solution is an example of the type of bespoke wrapper code that can be created using existing tools, such as the metapackages 'metaverse' and 'tidyverse'. Note that though brief, the syntax of each of the tools may take time to familiarise oneself with. This function accepts a dataframe, and splits multiple

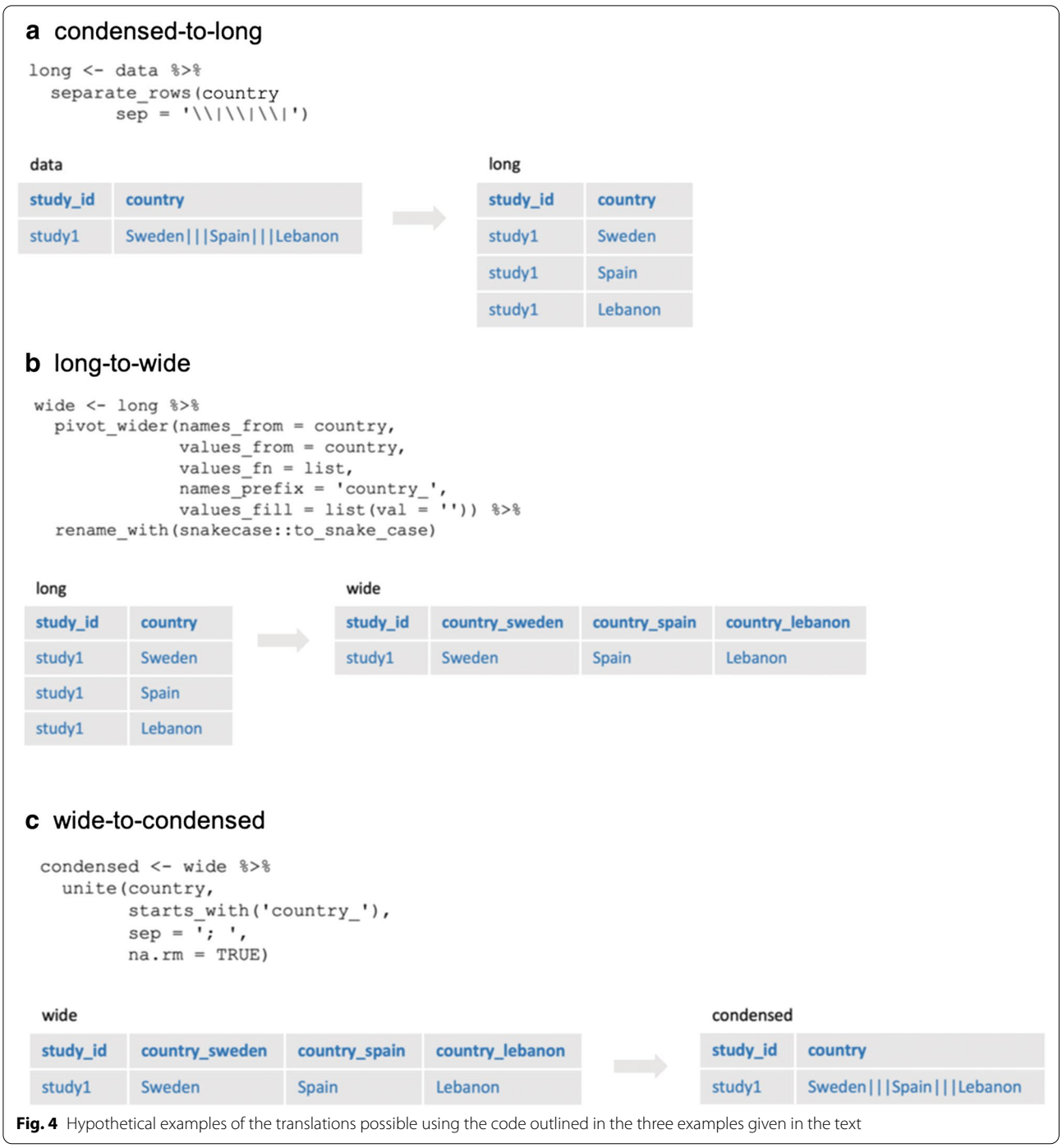

Fig. 4 Hypothetical examples of the translations possible using the code outlined in the three examples given in the text

values (separated by '|||') within a given column into rows, duplicating all other information in the dataframe for that row. The backslashes are necessary 'escape characters' to tell R that '|' does not have a purpose other than being a character in the string. See Fig. 4a for a hypothetical run-through.

```
long <- data %>%
  separate_rows(column1,
                sep = '\\|\\|\\|')
```

Example 2 Pivot from one row per level of a variable to one column per level of a variable (long-to-wide)

This function takes the long version of the database and converts each unique level of a variable into a column, and populates the resulting columns for rows with corresponding codes. It makes use of the pivot_wider() function from the 'tidyr' package [30]. The resultant database has one line per study, and each column is prefixed by the original column name for clarity. The to_snake_case() function from the 'snakecase' package [31] converts the column names into snake_case (i.e. words separated by underscores). See Fig. 4b for a hypothetical run-through.

```
wide <- long %>%
  pivot_wider(names_from = column1,
              values_from = column1,
              values_fn = list,
              names_prefix = 'column1_',
              values_fill = list(val = '')) %>%
  rename_with(snakecase::to_snake_case)
```

Example 3 Compressing multiple columns into a condensed format (wide-to-condensed)

The following function makes use of the 'tidyr' package [30] function unite() to take a wide database, consisting of one column per value of a variable, with different variables indicated by a column name prefix (e.g. 'column1_'), and produces a condensed table, with multiple values per cell, separated by ';' (demonstrating that the user can easily alter the delimiting separator within a translation process). The output retains the same number of rows as the wide dataframe. See Fig. 4c for a hypothetical run-through.

```
condensed <- wide %>%
  unite(column1,
        starts_with('column1_'),
        sep = '; ',
        na.rm = TRUE)
```

Documentation and resources (data dictionary)

Whatever choices evidence synthesists make in designing their databases, they should structure their data for use by other researchers (and their future selves). They should document this structure with a data dictionary, by stating what each column in a table refers to, and what each row defines. Synthesists should not be afraid to provide multiple tables, that is, a collection of tables. They should provide details as to what information each table provides, and how they connect to each other (which variables are common or *primary keys*). As the evidence synthesis community is still relatively unaware of data structures, consider providing readers with at least one resource on data structures; for example, entry-level texts such as '*R Packages*' [32] and the seminal 'Tidy Data' [33] may be appropriate.

Conclusions

It is clear from the literature that Open Data and Open Synthesis bring a range of benefits to the research community and evidence users [7, 34–39]. The way in which we secure this Open Synthesis future, however, is up for debate. Our tools aim to facilitate a pathway to a clear and easy future for Open Syntheses by establishing standard practices in the design and publication of databases of evidence produced within systematic review and map projects. We summarise our key recommendations in Table 3.

We appreciate that changing practices across a community will not be an easy task, with a potentially steep learning curve, issues around a lack of awareness, and resistance to change. However, we believe that these obstacles can be easily overcome by providing templates and tools that allow users to design, use and publish databases with ease (https://softloud.github.io/sysrevdata/). Indeed, by embracing standard approaches to database design, we believe that the job of producing and using databases can be made far easier through automation tools (e.g. using EviAtlas; [40]). We eagerly anticipate future toolchain walkthroughs that provide subdiscipline- or tool-specific code for structuring data for evidence synthesis.

Future work is needed to develop resources to support awareness raising and adoption of these tidy data

Table 3 Key recommendations for multifunctional, machine readable systematic review and map databases

Objective	Description
1. Select the right database format	Carefully plan what format databases are the most appropriate for data extraction, visualisation, and analysis
2. Ensure design allows translation	Select a database format and population method that allow translation between wide, long and condensed formats. Avoid condensing information where a loss of information or data linkage occurs
3. Choose an appropriate naming convention	Particularly for wide formats, use machine readable column names for related variables to facilitate aggregation. Populate cells with actual data labels rather than using binary labels ('0' and '1') to facilitate aggregation
4. Publish the highest resolution database	Provide the database with the highest level of resolution and information as a digital, machine readable additional file, from which condensed or other formats can be translated
5. Document your data	Include a detailed data dictionary or meta-data file/worksheet is included in the database file that fully describes each variable/column and coding, that is self-sufficient without the need to read the manuscript methods to understand

principles, and efforts are underway to produce these. Simultaneously, digital tools are needed to allow synthesists to immediately convert between human- and machine-readable formats, and between different machine-readable formats (i.e. long to wide) for different purposes. In addition, systematic review and map databases should be static and unchanged unless accompanied by clear methodological justification and reporting via a registered and peer-reviewed protocol and final report: these are the central principles of rigorous evidence synthesis. However, the advent of living systematic reviews [41] (and maps) implies the need for guidance and procedures on how databases should be updated transparently and reliably. Clear versioning may be the solution, but this is outside our intended scope and should be tackled by the living evidence synthesis community.

However, these tools will only be useful if the community of evidence synthesists embraces the need for consistency, the need for Open Data and Open Synthesis and the benefits for tidy systematic review and map databases. In doing so, we can maximise the efficiency of the conduct, updating and upgrading of evidence syntheses.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13750-021-00219-x.

Additional file 1. Database of systematic reviews and maps and descriptive information about their database formats. https://doi.org/10.6084/m9.figshare.13019186.

Acknowledgements
The authors wish to thank the funders and organisers of the Evidence Synthesis Hackathon 2018 (Mistra EviEM and the Fenner School of the Australian National University) for facilitating this collaboration between participants since the ESH in April 2018.

Authors' contributions
All authors developed the concept and drafted the manuscript. All authors developed the code. All authors read and approved the final manuscript.

Funding
Open access funding provided by Stockholm University. Contributions from NRH were covered by funding from the Alexander von Humboldt Foundation (Humboldt Research Fellowship for Experienced Researchers). All other contributions were provided on a voluntary basis.

Availability of data and materials
All data are available in supplementary information archived on figshare.com (https://doi.org/10.6084/m9.figshare.13019186).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Stockholm Environment Institute, Linnégatan 87D, Stockholm, Sweden. [2] Mercator Research Institute on Global Commons and Climate Change, Berlin, Germany. [3] Africa Centre for Evidence, University of Johannesburg, Johannesburg, South Africa. [4] School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. [5] Norwegian Institute for Nature Research, Torgarden, Postbox 5685, 7485 Trondheim, Norway.

Received: 6 October 2020 Accepted: 9 February 2021
Published online: 27 February 2021

References

1. Collaboration for Environmental Evidence. Guidelines and standards for evidence synthesis in environmental management, version 5.0 [Pullin AS, Frampton GK, Livoreil B, Petrokofsky G (eds)]. 2018. www.environmentalevidence.org/information-for-authors. Accessed 1 Dec 2020
2. Kohl C, McIntosh EJ, Unger S, Haddaway NR, Kecke S, Schiemann J, et al. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. Environ Evid. 2018;7(1):8.
3. Wolffe TA, Whaley P, Halsall C, Rooney AA, Walker VR. Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. Environ Int. 2019;130:104871.
4. Friesen J. Introducing JSON. Java XML and JSON. New York : Springer; 2019. p. 187–203.

5. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
6. Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, et al. The TRUST principles for digital repositories. Sci Data. 2020;7(1):144.
7. Haddaway NR. Open Synthesis: on the need for evidence synthesis to embrace Open Science. Environ Evid. 2018;7(1):26.
8. Vicente-Sáez R, Martínez-Fuentes C. Open Science now: a systematic literature review for an integrated definition. J Bus Res. 2018;88:428–36.
9. Gewin V. Data sharing: an open mind on open data. Nature. 2016;529(7584):117–9.
10. Stodden V, Miguez S. Best practices for computational science: software infrastructure and environments for reproducible and extensible research. Available at SSRN 2322276. 2013.
11. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK. Good enough practices in scientific computing. PLoS Comput Biol. 2017;13(6):e1005510.
12. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015;348(6242):1422–5.
13. Grainger MJ, Bolam FC, Stewart GB, Nilsen EB. Evidence synthesis for tackling research waste. Nat Ecol Evol. 2020;4(4):495–7.
14. Kostkova P, editor. A roadmap to integrated digital public health surveillance: the vision and the challenges. In: Proceedings of the 22nd international conference on World Wide Web; 2013.
15. Whitlock MC. Data archiving in ecology and evolution: best practices. Trends Ecol Evol. 2011;26(2):61–5.
16. Ioannidis JP, Khoury MJ. Improving validation practices in "omics" research. Science. 2011;334(6060):1230–2.
17. Fan B, Zhang X, Fan W, editors. Identifying physician fraud in healthcare with Open Data. In: International conference on smart health. New York: Springer; 2019.
18. Chan A-W, Song F, Vickers A, Jefferson T, Dickersin K, Gøtzsche PC, et al. Increasing value and reducing waste: addressing inaccessible research. Lancet. 2014;383(9913):257–66.
19. Sivarajah U, Weerakkody V, Waller P, Lee H, Irani Z, Choi Y, et al. The role of e-participation and open data in evidence-based policy decision making in local government. J Organ Comput Electron Commer. 2016;26(1–2):64–79.
20. Groom Q, Weatherdon L, Geijzendorffer IR. Is citizen science an open science in the case of biodiversity observations? J Appl Ecol. 2017;54(2):612–7.
21. Haddaway NR, Verhoeven JT. Poor methodological detail precludes experimental repeatability and hampers synthesis in ecology. Ecol Evol. 2015;5(19):4451–4.
22. Wolffe TA, Vidler J, Halsall C, Hunt N, Whaley P. A survey of systematic evidence mapping practice and the case for knowledge graphs in environmental health and toxicology. Toxicol Sci. 2020;175(1):35–49.
23. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.
24. Ojanen M, Zhou W, Miller DC, Nieto SH, Mshale B, Petrokofsky G. What are the environmental impacts of property rights regimes in forests, fisheries and rangelands? Environ Evid. 2017;6(1):12.
25. Bayliss HR, Schindler S, Adam M, Essl F, Rabitsch W. Evidence for changes in the occurrence, frequency or severity of human health impacts resulting from exposure to alien species in Europe: a systematic map. Environ Evid. 2017;6(1):21.
26. Sola P, Cerutti PO, Zhou W, Gautier D, Iiyama M, Schure J, et al. The environmental, socioeconomic, and health impacts of woodfuel value chains in Sub-Saharan Africa: a systematic map. Environ Evid. 2017;6(1):4.
27. Jones-Hughes T, Peters J, Whear R, Cooper C, Evans H, Depledge M, et al. Are interventions to reduce the impact of arsenic contamination of groundwater on human health in developing countries effective? A systematic review. Environ Evid. 2013;2(1):11.
28. Bååth R. The state of naming conventions in R. R J. 2012;4(2):74–5.
29. Grolemund G, Wickham H. R for data science. 2018.
30. Wickham H, Henry L. Tidyr: Tidy messy data. R package version 11. 2020.
31. Grosser M. Snakecase: convert strings into any case. R package version 0.11.0. 2019.
32. Wickham H. R packages: organize, test, document, and share your code. Newton: O'Reilly Media, Inc.; 2015.
33. Wickham H. Tidy data. J Stat Softw. 2014;59(10):1–23.
34. Arza V, Fressoli M. Systematizing benefits of open science practices. Inf Serv Use. 2017;37(4):463–74.
35. Nielsen M. Reinventing discovery: the new era of networked science. Princeton: Princeton University Press; 2020.
36. David PA. The economic logic of "open science" and the balance between private property rights and the public domain in scientific data and information: a primer. The role of the public domain in scientific and technical data and information. Stanford: Stanford Inst. for Economic Policy Research; 2003. p. 19–34.
37. Hartshorne J, Schachner A. Tracking replicability as a method of post-publication open evaluation. Front Comput Neurosci. 2012;6:8.
38. Surowiecki J. The wisdom of crowds. New York: Anchor; 2005.
39. Wiggins A, Crowston K, editors. From conservation to crowdsourcing: a typology of citizen science. In: 2011 44th Hawaii international conference on system sciences. New York: IEEE; 2011.
40. Haddaway NR, Feierman A, Grainger MJ, Gray CT, Tanriver-Ayder E, Dhaubanjar S, et al. EviAtlas: a tool for visualising evidence synthesis databases. Environ Evid. 2019;8(1):1–10.
41. Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, et al. Living systematic review: 1. Introduction—the why, what, when, and how. J Clin Epidemiol. 2017;91:23–30.

Publisher's Note