Hierarchical Data Extraction

A New Approach to Capturing Repeating Data Sets in Systematic Reviews

By Peter O'Blenis, January 2017

Data extraction, or abstraction, is the process of reading through a report of a clinical study and extracting all of the key bits of information that you wish to analyze as part of your systematic review.

Some extracted data is simple and only occurs once in a paper - the study type or the therapeutic area being examined, for example.

Other types of data may be repeated throughout a paper. Studies may examine multiple outcomes, and each outcome may have multiple study arms, with results for each study arm measured across multiple time periods. The number of each of these data subsets can vary widely from study to study. **(See Figure 1).**

Essentially, data to be extracted can be of arbitrary breadth (the number of different subsets of data) and depth (the number of layers), with relationships between different subsets that must be preserved for the context of the data to remain intact.

A long-standing challenge in the systematic review community has been how to design a data extraction form, or forms, to capture related, repeating blocks of information.

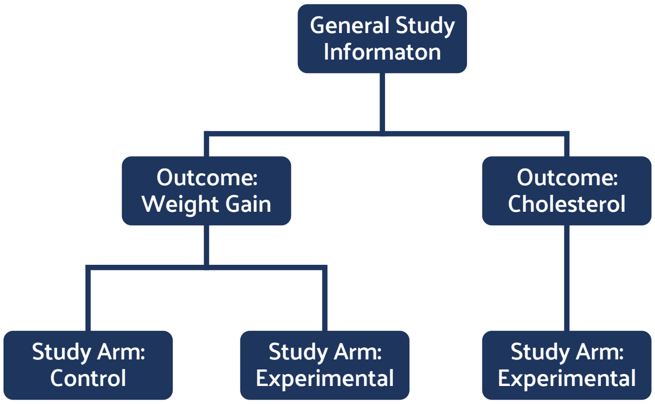


Figure 1: The chart illustrates a subset of a typical information hierarchy found in a research paper.

The simplest and most common approach has traditionally been to extract the data directly into a spreadsheet, inserting new rows and columns as relevant data is found in the papers. Unfortunately, spreadsheets lack many of the valuable qualities of electronic forms: they are not multi-user, they don't typically have version control or audit trails, they are subject to being lost, deleted or overwritten, and they do not validate input at the point of entry.

“There are many reasons why spreadsheets are the wrong tool for data extraction, but in the face of complex data sets, it is easy to understand why they are so broadly used.”

Another common response to the challenge of repeating data is to design data extraction forms that try to anticipate the maximum number of each possible dataset, and simply repeat question sets within the form. This approach has two main drawbacks:

1. The forms can get very large as the form designer attempts to anticipate all contingencies; and,
2. Data exported from these forms is flat, rather than hierarchical, making the exported data very difficult to work with (more on this later).

Breaking the Form Up

To get around the issue of giant data extraction forms with large sections of unused repeating question sets, you might decide to break your forms up into repeating and non-repeating sections.

Non-repeating data, such as study characteristics or Risk of Bias assessment, can be completed once on a standard form.

Repeating data, such as study arms, outcomes and time periods, can be captured on separate forms where a new copy of each form is used for each repeating data group (e.g. if you have four time periods, you would complete four copies of the ‘time periods’ data extraction form).

This solution allows the reviewer to instantiate just the data extraction forms that they need for any given study and, because exactly the same questions are used for each instance of the data, the captured data can output vertically in table form, making it easy to work with.

While this approach goes a long way towards addressing the challenges of complex data extraction, it does have a couple of significant drawbacks.

First, repeating forms can result in a lot of small forms being generated. Navigating this stack of forms (even electronic forms) can present a significant challenge to the reviewer and the project coordinators.

What’s more, the relationships between each form must be tracked to preserve the integrity of the data (e.g. each time period must be associated with an outcome, each outcome must be associated with a study arm, etc). Maintaining and tracking relationships adds overhead to the review process and introduces the possibility of error.

Hierarchical Data Extraction

Hierarchical Data Extraction (HDE) was designed to be a more elegant solution for dealing with repeating data sets. HDE is based on the following key concepts behind repeating forms:

- A form is constructed for each type of repeating data.
- Users can instantiate these forms on an as-needed basis as they find new information in a paper.

HDE extends the repeating form approach in three significant ways:

- Forms are associated with one another through parent-child relationships.
- The software automatically creates and maintains these relationships in the background as the user instantiates new forms.
- The collection of instantiated forms is navigable using a tree structure with individual form instances as nodes.

Configuring HDE

One of the benefits of HDE is its flexibility and ease of configuration. With the[1] following simple steps, you can set up HDE in your systematic review project :

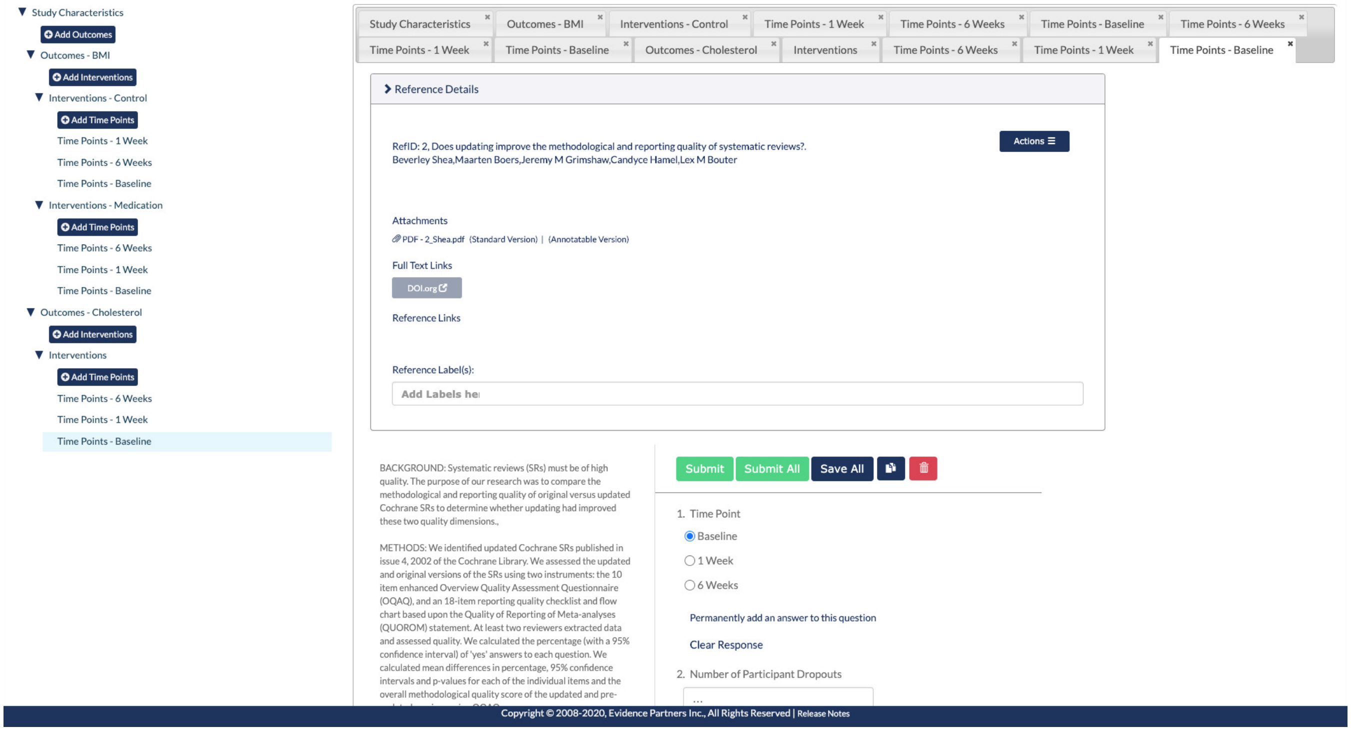
1. Create a root parent form to capture the data at the top of your information hierarchy. Root forms may be single instance forms, such as Study Characteristics, or multiple instance, such as Time Periods.

2. Create forms to capture each type of repeating dataset you expect to find and report on in your review.

3. Link all forms except the root form to its logical parent. (e.g. Time Period links to Study Arm as its parent since you can have multiple Time Periods in a Study Arm).

That’s it. You have now configured a data extraction instrument that can handle an arbitrary number of repeating data sets at any level of your information hierarchy.

[1] HDE is fully supported in Distiller systematic review software. Other software tools may not offer this functionality.



RefID: 2, Does updating improve the methodological and reporting quality of systematic reviews?.
Beverley Shea,Maarten Boers,Jeremy M Grimshaw,Candyce Hamel,Lex M Bouter

Attachments

PDF - 2_Shea.pdf (Standard Version) | (Annotatable Version)

Full Text Links

Reference Links

Reference Label(s):

BACKGROUND: Systematic reviews (SRs) must be of high quality. The purpose of our research was to compare the methodological and reporting quality of original versus updated Cochrane SRs to determine whether updating had improved these two quality dimensions.,

METHODS: We identified updated Cochrane SRs published in issue 4, 2002 of the Cochrane Library. We assessed the updated and original versions of the SRs using two instruments: the 10 item enhanced Overview Quality Assessment Questionnaire (OQAQ), and an 18-item reporting quality checklist and flow chart based upon the Quality of Reporting of Meta-analyses (QUOROM) statement. At least two reviewers extracted data and assessed quality. We calculated the percentage (with a 95% confidence interval) of 'yes' answers to each question. We calculated mean differences in percentage, 95% confidence intervals and p-values for each of the individual items and the overall methodological quality score of the updated and pre-

1. Time Point

● Baseline

○ 1 Week

○ 6 Weeks

Permanently add an answer to this question

Clear Response

2. Number of Participant Dropouts

...

Copyright © 2008-2020, Evidence Partners Inc., All Rights Reserved

Figure 2: Figure description

Reviewing Using HDE

As noted earlier, one of the challenges with repeating forms is that reviewers often need to instantiate a lot of them for any given study. This poses challenges in tracking and navigation as reviewers move from one form to another, filling in data as they find it.

To help manage forms and to make navigation of instantiated forms intuitive, HDE uses a dynamic navigation tree.

When a reviewer first starts reviewing a paper, only the root form is available in the tree. As the reviewer finds repeating data sets, children can be added to the root node. **(Figure 2)**

For example, as reviewers find outcomes, they can add an Outcomes form for each one. Study arms can then be added as children to the appropriate Outcome, and so on.

In this way, the reviewer is instantiating the forms that are needed, and constructing the information hierarchy specific to each specific paper being reviewed, on the fly. The reviewers can then navigate between forms by simply clicking the tree node for the desired form.

One Last "Key" to HDE

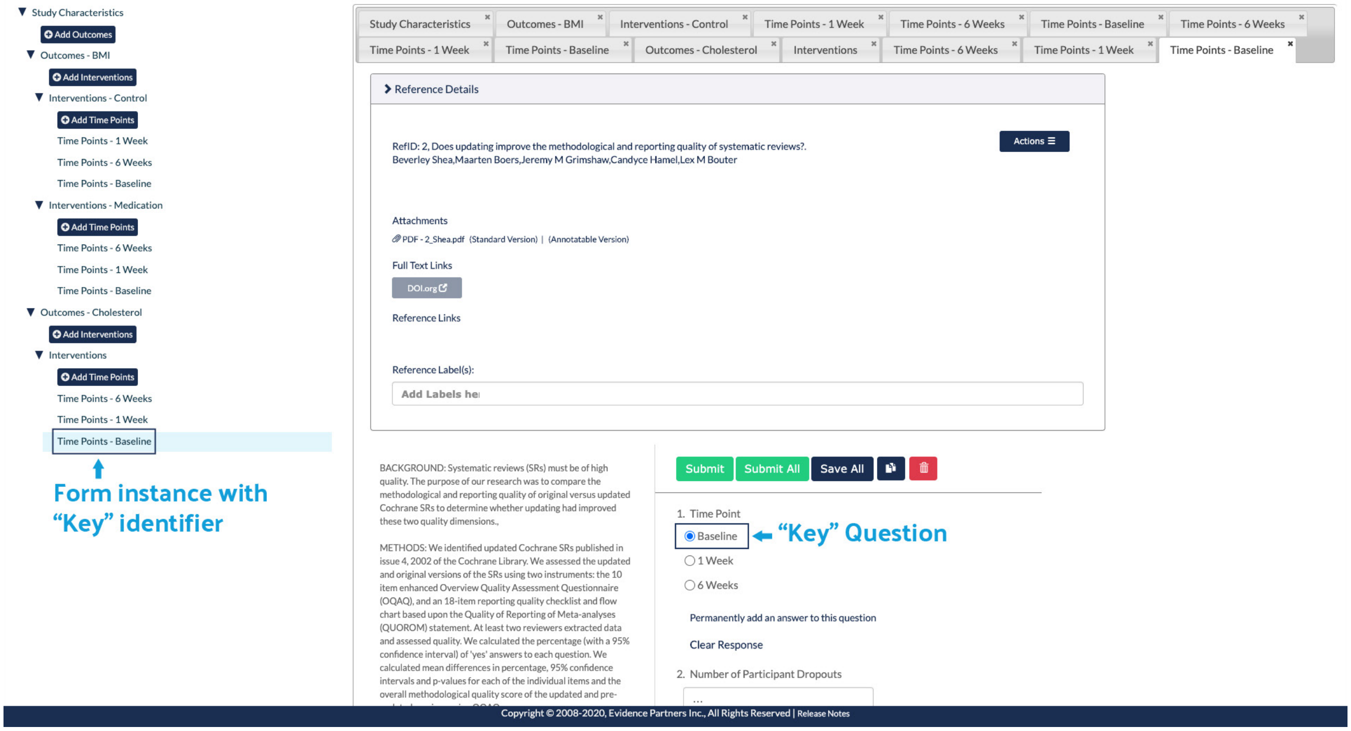
As we discussed, the software automatically tracks parent-child relationships between forms as the reviewer instantiates new repeating forms. That said, if the nodes in the navigation tree only contained the form names without any additional information, it would be difficult for the reviewer to differentiate between them.

For example, let's say an outcome is measured across three time periods: 1 Week, 2 Week and 4 Week. After filling out a form for each time period, how would the reviewer know, by looking at the navigation tree, which repeating Time Period form contained data for a specific time period? They would likely need to click through them until they found the one they needed.

HDE addresses this by allowing the user, at configuration time, to designate a question, or group of questions, on each form as a unique identifier, or "key" for that specific form.

The time period question (e.g. Time since intervention = x weeks) on a Time Period form could be used as that form's key. Once the user answers a key question on a form, that question's response is appended to the form name in the navigation tree.

This allows to user, at a glance, to see exactly which instances of a repeating form have been created within each section of the tree, making navigation between forms intuitive. **(See Figure 3)**



- Study Characteristics
  - Add Outcomes
  - Outcomes - BMI
    - Add Interventions
    - Interventions - Control
      - Add Time Points
      - Time Points - 1 Week
      - Time Points - 6 Weeks
      - Time Points - Baseline
    - Interventions - Medication
      - Add Time Points
      - Time Points - 6 Weeks
      - Time Points - 1 Week
      - Time Points - Baseline
  - Outcomes - Cholesterol
    - Add Interventions
    - Interventions
      - Add Time Points
      - Time Points - 6 Weeks
      - Time Points - 1 Week
      - Time Points - Baseline

Form instance with “Key” identifier

Study Characteristics × | Outcomes - BMI × | Interventions - Control × | Time Points - 1 Week × | Time Points - 6 Weeks × | Time Points - Baseline × | Time Points - 6 Weeks ×

Time Points - 1 Week × | Time Points - Baseline × | Outcomes - Cholesterol × | Interventions × | Time Points - 6 Weeks × | Time Points - 1 Week × | Time Points - Baseline ×

Reference Details

RefID: 2, Does updating improve the methodological and reporting quality of systematic reviews?.
Beverley Shea,Maarten Boers,Jeremy M Grimshaw,Candyce Hamel,Lex M Bouter

Actions ☰

Attachments

PDF - 2_Shea.pdf (Standard Version) | (Annotatable Version)

Full Text Links

DOI.org

Reference Links

Reference Label(s):

Add Labels he

BACKGROUND: Systematic reviews (SRs) must be of high quality. The purpose of our research was to compare the methodological and reporting quality of original versus updated Cochrane SRs to determine whether updating had improved these two quality dimensions.,

METHODS: We identified updated Cochrane SRs published in issue 4, 2002 of the Cochrane Library. We assessed the updated and original versions of the SRs using two instruments: the 10 item enhanced Overview Quality Assessment Questionnaire (OQAQ), and an 18-item reporting quality checklist and flow chart based upon the Quality of Reporting of Meta-analyses (QUOROM) statement. At least two reviewers extracted data and assessed quality. We calculated the percentage (with a 95% confidence interval) of 'yes' answers to each question. We calculated mean differences in percentage, 95% confidence intervals and p-values for each of the individual items and the overall methodological quality score of the updated and pre-

1. Time Point

● Baseline ← “Key” Question

○ 1 Week

○ 6 Weeks

Permanently add an answer to this question

Clear Response

2. Number of Participant Dropouts

...

Copyright © 2008-2020, Evidence Partners Inc., All Rights Reserved | Release Notes

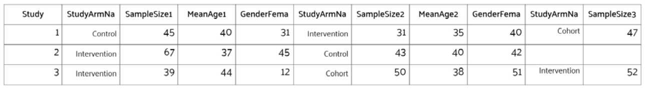
Figure 3: Figure description

Generating Output

As most experienced reviewers will tell you, capturing data is only half the battle. Aggregating, collating and getting the data into a useful structure for presentation or analysis can also be a significant challenge.

Repeating forms were originally developed to assist with the challenges of capturing and reporting on repeating data sets. The major advancement of repeating forms was the ability to generate tables of like data (e.g. Time Points, Study Arms, Interventions, etc) easily, rather than having different instances of repeating data presented as a single horizontal table that requires significant manual manipulation to work with.

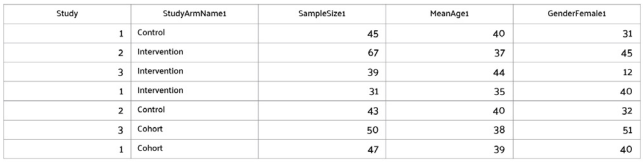
Below is an example of a recurring Study Arm data captured horizontally **(Figure 4)**. This data will be difficult to work with and will likely require manual reconfiguration before it can be analyzed.



Study	StudyArmNa	SampleSize1	MeanAge1	GenderFema	StudyArmNa	SampleSize2	MeanAge2	GenderFema	StudyArmNa	SampleSize3
1	Control	45	40	31	Intervention	31	35	40	Cohort	47
2	Intervention	67	37	45	Control	43	40	42		
3	Intervention	39	44	12	Cohort	50	38	51	Intervention	52

Figure 4: Figure description

Now consider this example **(Figure 5)** of a recurring Study Arm data reported vertically:



Study	StudyArmName1	SampleSize1	MeanAge1	GenderFemale1
1	Control	45	40	31
2	Intervention	67	37	45
3	Intervention	39	44	12
1	Intervention	31	35	40
2	Control	43	40	32
3	Cohort	50	38	51
1	Cohort	47	39	40

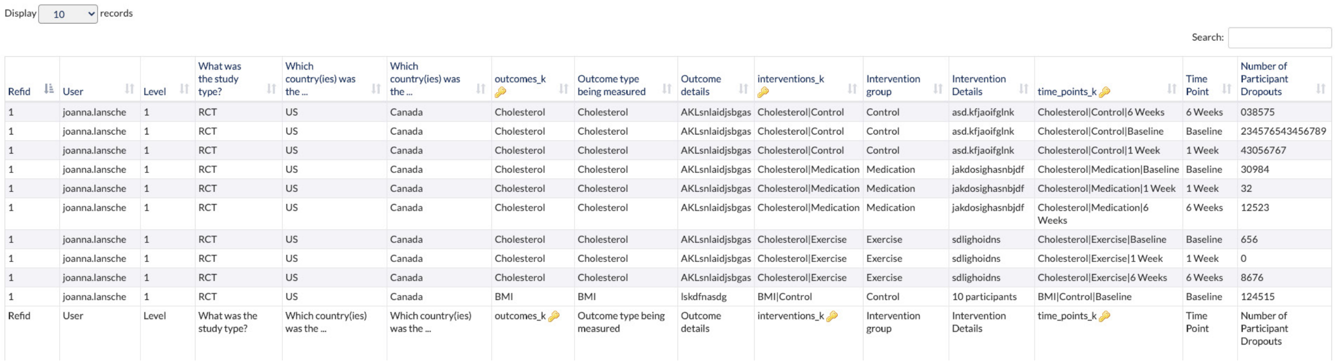
Figure 5: Figure description

Because HDE is based on the repeating forms concept, it can provide a vertical representation of any data captured at any level of its information hierarchy - but it actually goes one important step beyond that.

Since HDE maintains parent-child relationships between all forms automatically, it can merge data from different levels of the information hierarchy and present them in a vertical or hierarchical format that is intuitive to read and analysis-ready.

In the output below **(Figure 6)**, we see data captured for reference paper number 11564.

- We see that Study Characteristics of Year, Trial Name and Study Design were captured once, using a single instance form. This data is repeated on each row of the report.
- We see that 2 outcomes were reported: BMI and Cholesterol.
- In the BMI Outcome, there were two Arms reported: Control and Experimental Therapy.
- In the BMI->Experimental Therapy arm we see both 1 Week and 2 Week follow ups under Time Period.



Display 10 records

Search:

Refid	User	Level	What was the study type?	Which country(ies) was the ...	Which country(ies) was the ...	outcomes_k 🔑	Outcome type being measured	Outcome details	interventions_k 🔑	Intervention group	Intervention Details	time_points_k 🔑	Time Point	Number of Participant Dropouts
1	joanna.lansche	1	RCT	US	Canada	Cholesterol	Cholesterol	AKLsnlaidjsbgas	Cholesterol|Control	Control	asd.kfjaoifglnk	Cholesterol|Control|6 Weeks	6 Weeks	038575
1	joanna.lansche	1	RCT	US	Canada	Cholesterol	Cholesterol	AKLsnlaidjsbgas	Cholesterol|Control	Control	asd.kfjaoifglnk	Cholesterol|Control|Baseline	Baseline	234576543456789
1	joanna.lansche	1	RCT	US	Canada	Cholesterol	Cholesterol	AKLsnlaidjsbgas	Cholesterol|Control	Control	asd.kfjaoifglnk	Cholesterol|Control|1 Week	1 Week	43056767
1	joanna.lansche	1	RCT	US	Canada	Cholesterol	Cholesterol	AKLsnlaidjsbgas	Cholesterol|Medication	Medication	jakdosighasnbjdf	Cholesterol|Medication|Baseline	Baseline	30984
1	joanna.lansche	1	RCT	US	Canada	Cholesterol	Cholesterol	AKLsnlaidjsbgas	Cholesterol|Medication	Medication	jakdosighasnbjdf	Cholesterol|Medication|1 Week	1 Week	32
1	joanna.lansche	1	RCT	US	Canada	Cholesterol	Cholesterol	AKLsnlaidjsbgas	Cholesterol|Medication	Medication	jakdosighasnbjdf	Cholesterol|Medication|6 Weeks	6 Weeks	12523
1	joanna.lansche	1	RCT	US	Canada	Cholesterol	Cholesterol	AKLsnlaidjsbgas	Cholesterol|Exercise	Exercise	sdlighoidns	Cholesterol|Exercise|Baseline	Baseline	656
1	joanna.lansche	1	RCT	US	Canada	Cholesterol	Cholesterol	AKLsnlaidjsbgas	Cholesterol|Exercise	Exercise	sdlighoidns	Cholesterol|Exercise|1 Week	1 Week	0
1	joanna.lansche	1	RCT	US	Canada	Cholesterol	Cholesterol	AKLsnlaidjsbgas	Cholesterol|Exercise	Exercise	sdlighoidns	Cholesterol|Exercise|6 Weeks	6 Weeks	8676
1	joanna.lansche	1	RCT	US	Canada	BMI	BMI	lskdfnasdg	BMI|Control	Control	10 participants	BMI|Control|Baseline	Baseline	124515
Refid	User	Level	What was the study type?	Which country(ies) was the ...	Which country(ies) was the ...	outcomes_k 🔑	Outcome type being measured	Outcome details	interventions_k 🔑	Intervention group	Intervention Details	time_points_k 🔑	Time Point	Number of Participant Dropouts

Figure 6: Figure description

> **HDE automatically maintains the critical information hierarchy and presents the data in a way that is easy to work with and manipulate to meet specific reporting requirements.**

Conclusion

Capturing and reporting on complex repeating data sets has been an ongoing challenge for the systematic review community for more than a decade. Typically, it has been addressed by creating unwieldy spreadsheets or bloated forms to handle all the possibilities within a dataset.

HDE was developed to provide a cleaner, more efficient way to deal with recurring, hierarchical data sets. It is intuitive to configure, easy for reviewers to navigate and it exports complex data into formats that are easy to work with.

Toll free: (844) 622-8727
505 March Road, Suite 450
Ottawa, Ontario, Canada, K2K 3A4
distillersr.com