Open Repair Data: Aggregation for FixFest 2019

This document is a summary of work undertaken to produce the first aggregation of partner data, along with recommendations for next steps to continue this work.

Overview

- All partner data mapped and combined into an aggregated dataset of near 30,000 records - https://openrepair.org/open-data/downloads/
- This is an initial aggregation to prove the concept, and has provided further insight into changes required for the next version of ORDS
- Fields related to **language** should be added to ORDS, and the **product category** and **repair status** fields require more analysis for consistent mapping
- Along with agreement on mappings, tool support will be required to facilitate aggregation on a regular basis

In September 2019, the Open Repair Alliance published our first aggregated set of Open Repair Data.

We timed this to coincide with Fixfest Berlin, a major date in the repair community calendar. For the first time at Fixfest, repair data was under a spotlight - see Opening Up Repair Data At Fixfest 2019.

Why are we aggregating repair data?

We know that citizen data on environmental issues can have direct influence at the policy level. As such, the Open Repair Alliance works on an open data standard for repair data, and ORA members work together to combine the repair data from our community repair events. By mapping this data to a common format, we can pool our repair data together and look for patterns and trends to help inform policy.

Work started two years ago to collaborate on an open repair data standard. Now, partners are all using tools in which data is collected electronically. We have now published a dataset of combining near 30,000 records of repair attempts recorded by partners.

Our first dive into the data was at a pre-Fixfest event held at Mozilla Berlin where around 20 of us used the newly-published data to investigate fault types. We've blogged about the event in our post Why Do Computers Fail? Insights From Fixfest 2019.

Where does the repair data come from?

Publication of the dataset was made possible by contributions from ORA partners, all of whom are collecting repair data at their community repair events. For the release for Fixfest, we had 15,914 records included, with this now at 29,229 records.

Partner	Data	Collection method
Anstiftung	3,517 records provided in CSV format with 2,938 published. Language: German Last repair date: 2019-09-01	Network Reparatur Initiativen tool
Fixit Clinic	493 records provided in Excel format with 343 published. Language: English Last repair date: 2019-08-20	Google Form
Repair Café Foundation	19,079 record provided in Excel format with 13,315 published. Language: Multiple (mostly Dutch) Last repair date: 2019-08-31	Repair Monitor
The Restart Project	12,633 records provided in CSV format with/ 12,633 published. Language: Multiple (mostly English) Last repair date: 2019-09-08	Fixometer in Restarters.net

More information can be found on the data downloads page.

How did we aggregate the data?

Each tool currently collects repair data in a slightly different format. Therefore, each data set needed to be mapped to the ORDS format.

This was a first run through of the aggregation and mapping, surfacing ideas for shared codelists and tool support to enable regular aggregation from partner datasets. Much of the work in this first run through was undertaken manually. Future work of the ORA is in refining these codelists, and enabling each tool to produce the data in the ORDS format by each partner, to be more easily combined.

Defining mappings

Each set of partner data was imported to a Google sheet and the sheet structure analysed for columns that could be mapped to ORDS values and to determine the required manipulations to do so. Some columns have relatively straight-forward mappings across partners, whereas others, such as product category and repair status, are the subject of further work.

- Translations were executed using Translate My Sheet, an add-on which uses Google Translation API.
- Some data cleaning was carried out, e.g. removing newline characters from text fields.
- Some values had to be extracted from other values, e.g. group identifiers embedded in URLs, dates embedded in identifiers, regular expressions were used to parse and format these values.
- The raw records held identifiers that were unique to the provider but not necessarily unique across all datasets therefore a unique_id was formulated using the ORDS import number concatenated with a provider identifier.

Product categories

In order to produce a first aggregated data set, and to surface questions to be explored further on product categorisation, we started with an existing product category list and mapped data sets to this.

There are various criteria associated with category lists that can balance the lists utility for policy versus the each of the data collection (always bearing in mind that this data is collected by volunteers at busy community repair events). We will be reporting in more depth on these criteria soon. Our starter category list uses a medium granularity of 35 product categories, which we felt provides for a good balance of utility and ease of collection. We will be reporting in more detail on category list criteria and existing category lists soon. The ORDS standard is electronics and electricals only, so the category list does not include electronics and electricals at present.

The initial ORDS category list:

product_category_ords
Aircon/Dehumidifier
Battery/charger/adapter
Decorative or safety lights
Desktop computer
Digital Compact Camera
DLSR / Video Camera
Fan
Flat screen
Hair & Beauty item
Handheld entertainment device
Headphones
Hi-Fi integrated
Hi-Fi separates
Kettle
Lamp
Laptop
Large home electrical
Misc
Mobile
Musical instrument
Paper shredder
PC Accessory
Portable radio
Power tool
Printer/scanner
Projector
Sewing machine
Small home electrical
Small kitchen item
Tablet
Toaster
Toy
TV and gaming-related accessories
Vacuum
Watch/clock

Only records that mapped to a non-”Misc” ORDS category are included in the ORDS export for Anstiftung, Fixit Clinic and Repair Cafe Foundation. Restart Project “Misc” items were included as it is assumed that repair parties deal with electrical items only.

See Appendix A for more detail on each of the existing partner members category lists, and how they map to the initial ORDS category list.

Repair status

Similar to product categories, each partner currently collects repair status information in their own way. Similarly, to enable aggregation, we have started with an initial list of statuses that balances between the use cases of the data and the ease of collection and mapping.

Initial ORDS repair status list
Fixed
Repairable
End of life
Unknown

The initial list is close to that already used by a number of the partners. This list of statuses can be refined following further work on requirements.

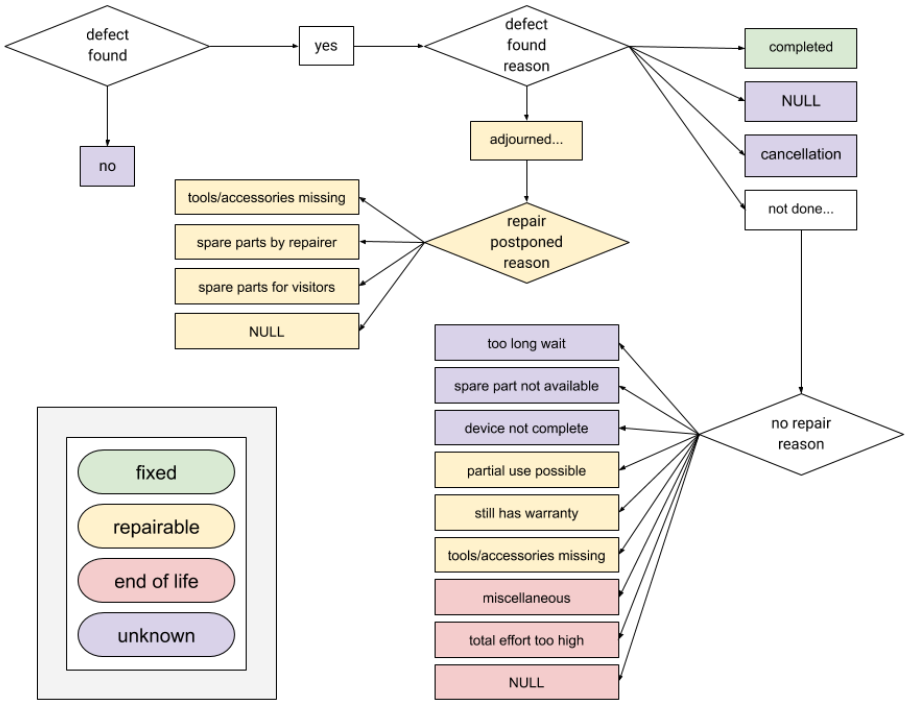
Restart Project

Original	**ORDS repair status**
Fixed	Fixed
Repairable	Repairable
End of life	End of life
Unknown	Unknown

The Restart Project has substatuses of Repairable and End-of-Life (including what the barriers to repair are), and also records whether spare parts were (or would be) required for the repair.

Anstiftung

Anstiftung data has a number of fields that determine the repair status. The below diagram shows the approach used for determining the mapping.



Fixit Clinic

The column “At this point: what's the disposition of this item?” on Fixit Clinic’s form represents the repair outcome of the repair attempt. It is completed by the participant themselves, after the event. Many of the values could be mapped to the ORDS `repair_status` and there were many more that contained unique free-text data, possibly historic, or possibly entered into the wrong column by mistake.

Original	**ORDS repair_status**
Fixed (hooray!)	Fixed
Repairable and I’m going to keep trying	Repairable
Unserviceable (End-of-Life)	End of life
Unknown	Unknown

Repair Cafe Foundation

Mapped from column “Gerepareerd, ja, half/advies, nee” (“Repaired, yes, half / advice, no”), containing 7 unique, white-space trimmed values. The 3 unmapped values appear to be data-entry errors.

Dutch	**English**	**ORDS Repair Status**
Nee	No	Unknown
Ja	Yes	Fixed

Half en/of advies gegeven	Half and / or advice given	Repairable
werkt niet	does not work	End of life
Ketting eraf	Chain off	
Doet het niet goed	Does not do well	
Snoer is niet goed meer	Cord is no longer good	

Repair status of of No has currently been mapped as ‘Unknown’ rather than End of life - with further clarification needed as to when ‘No’ can be mapped specifically to End of life.

What were the results?

The first export of the aggregated partner data can be downloaded from the Open Repair Alliance data downloads page.

What next?

This initial mapping exercise has provided an excellent starting point for continuing data aggregation, and has highlighted a number of areas requiring further work in order to enable regular aggregation.

ORDS vNext

In addition to work on product categories and repair status, several other recommended (and more straightforward!) updates to the standard have been identified.

Refinements

Product category and repair status

As part of this work we have identified a number of common criteria in product category lists, and in repair statuses. We have started work to examine these criteria in more depth, and to see how they map to the use cases of ORDS data. Once this is done partners can discuss how to match up use cases with ease of collection, and recommendations can be made as to the canonical ORDS lists for product categories and repair statuses.

Multi-lingual support

Given that the ORA partners are language-diverse, in order to map as accurately as possible the options should be available in a variety of languages. To start with, the Google translations of partner categories should be reviewed by them.

ORA partners to be consulted on whether the provision of translated ORDS categories would be beneficial.

Future reference values need to be agreed upon by all partners with regard to language and meaning.

Additions

Data provider name

Which partner organisation was the data collected by. This is easy to add in and allows for filtering and reporting by provider of the aggregated data set.

Unique identifier

A convention for provision of an identifier that is unique to each record in the merged dataset is required. Experimentally the format [provider_id]_[unique_number] has been used although a `provider` field and unique `id` field is probably preferable.

Language

Dataset files may be supplied in originating language.

The data export and visualisations show partner data transformed with English `product_category` and `repair_status` as well as column/field headings in English. The `problem_text` values remain in the original language.

ORDS mapping should consider translations of the mapped data.

Download pages should include language information.

Country of origin

In which country was the data collected. This allows for filtering and reporting of the aggregated data set by country. A number of partners collect data from repair groups in various countries, so this can't be determined from provider alone - it needs to be recorded by the partner organisation, most likely by knowing the country in which the group that collected the data is based.

ORDS could be updated to include an ISO `country` field to hold this information.

Removals

Model has a low level of collection across partners. It could be removed as a required field from the standard.

Data Quality

Some of the data mapped quite well, core information was abundant and was useful for reporting purposes, e.g. date, product category, repair status

Some core fields contained scant or poor quality data, e.g. `model`, `year_of_manufacture`

field = value	**% of total**
repair_status = "Unknown"	4.06%
product_category = "Misc"	10.17%
problem = "" (empty/blank)	19.92%
brand = "Unknown"	55.85%
model = "Unknown"	72.35%
year_of_manufacturer = "????"	92.32%

Further work is needed both in assessing why certain fields have low data quality, and can be followed by work on improving support to repair groups to help mitigate these issues.

Tool support and process improvements for regular aggregation

For this initial aggregation of data, much of the work has been done manually. This is not sustainable for regular aggregation. Following agreement on lists for categories and repair statuses, mapping to an aggregated data set will become easier. It may require data mapping tools to be created to support automatic mapping, or at least simplify manual mapping.

Additionally, some further steps could be taken to streamline aggregation. Cleaning of data and formatting by providers at source would mean basic data cleaning doesn't need to be done during aggregation. If providers published data automatically to a registered location, data sets could be pulled together automatically.

Appendices

A: Existing product category lists

Restart Project

- 36 unique categories selected from a list at the point of data entry.
- Mapped manually to the ORDS product categories.
- Occasional flaws in the original categorisation e.g. “Electric Other ~ Camera, analog”.
- “Misc” items were included as it is assumed that repair parties deal with electrical items only.

Anstiftung

- 270 unique product categories
- The values seem to be a concatenation of selected options and free-text, e.g. “Computer ~ Laptop” and “Elektro Sonstiges ~ Katzenklappe” (“Electric Other ~ cat flap”).
- Mapped manually to the ORDS product categories.
- Only records that mapped to a non-Misc ORDS category were included in the ORDS export.
- Occasional flaws in the original category/kind relations e.g. “Electric Other ~ Camera, analog”.
- Some of the category/kind relations appear to be lost in Google translation, e.g. the variable translation of the word “Haushaltsgeräte” as shown below.
- Top level category keyword “Appliances” seems to cover both electrical and non-electrical.

Category	**Google translation**
Haushaltsgeräte ~ Backautomat	Appliances ~ Bakeware
Haushaltsgeräte ~ Standmixer	Appliances ~ Blender
Haushaltsgeräte ~ Kerzenhalter	Appliances ~ Candlesticks
Haushaltsgeräte ~ Kettensäge	Appliances ~ Chainsaw
Haushaltsgeräte ~ Strickmaschine	Home Appliances ~ Knitting Machine
Haushaltsgeräte ~ Postkartenautomat	Home appliances ~ Postcard machine
Haushaltsgeräte ~ Brotschneidemaschine	Household appliances ~ bread slicer
Haushaltsgeräte ~ Stuhl	Household appliances ~ chair
Haushaltsgeräte ~ Kaffeemühle	Household appliances ~ coffee grinder

Haushaltsgeräte ~ Besteck	Household appliances ~ cutlery

Fixit Clinic

- 329 unique categories.
- The values seem to be entered as free-text without any validation.
- Mapped manually to the ORDS product categories.
- Only records that mapped to a non-Misc ORDS category were included in the ORDS export.

Repair Cafe Foundation

- Concatenation of columns “Category” and “Kind of product”.
- Contained 1215 unique, white-space trimmed values.
- The values seem to be a concatenation of selected options and free-text, e.g. “Computer ~ Laptop” and “Elektro Sonstiges ~ Katzenklappe” (“Electric Other ~ cat flap”).
- Mapped manually to the ORDS product categories.
- The top-level categories are not mappable to ORDS without considering “Kind of product”.

Partners are currently looking at other product and category schemas, and other taxonomies of products, such as iFixit, to assess how they could fit the requirements of the open repair data standard.