OpenActive Data Quality Reporting Framework

Version 0.3 - December 2022

Introduction

Publishing open data about where, when and what activities take place can make it easier for people to find and book activities online. OpenActive provides a standard format to publish and use data about sport and physical activities.

Now in its sixth year, OpenActive has grown significantly in the range of data publishers and in the number and variety of opportunities for sport and physical activity. As the initiative scales and approaches maturity, this is an ideal time to broaden the focus from 'opening up data' to exploring the whole experience of creating, sharing and using OpenActive data to create value, both economic and social.

Feedback from data users has highlighted areas where the end user experience when searching for activities could be improved. Additionally, as this phase of the initiative aligns more with Sport England's Uniting the Movement strategy, we need to understand better where OpenActive data could support new use cases. This framework outlines our approach to data quality (DQ) across the initiative.

Approach

Our DQ reporting framework is adapted from the Office of National Statistics data quality action planning approach. This focuses attention on the purposes the data is being used for. In OpenActive, we often refer to these purposes as 'use cases' - the different ways that the OpenActive standards and specifications can be used to solve a problem or to take advantage of an opportunity to help enable more people to be active.

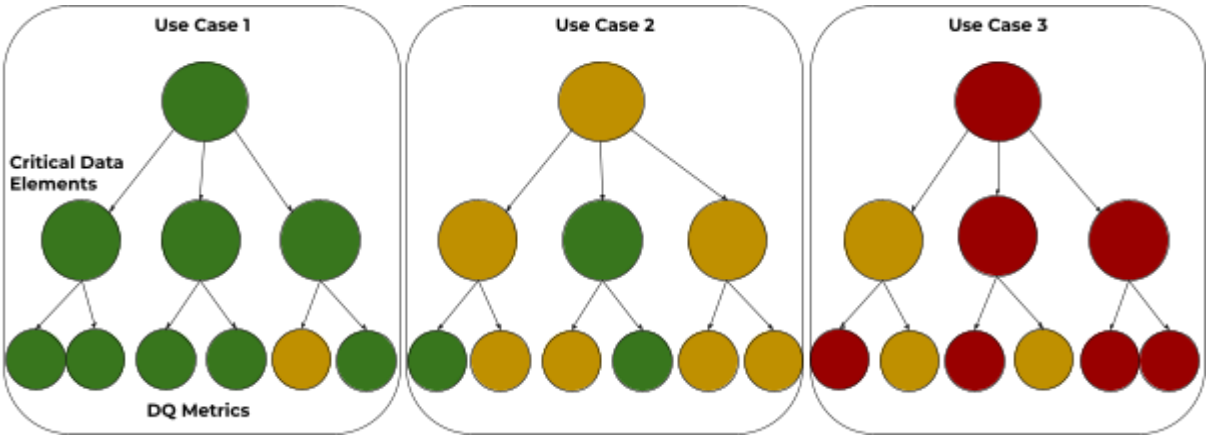
For each use case, we:

1. Consider what you need from the data - identify the *critical data elements*
2. Consider what 'good quality' looks like
3. Develop some measures or metrics to describe the current state
4. Report on these measures on a regular basis
5. Refer to these to identify and drive actions to improve data quality in ways that will directly support the purpose

Over recent meetings of the OpenActive W3C community group, we have considered steps 1 to 3 above.

Our proposal is to develop and publish a dashboard to display a set of data quality metrics that describe the state of OpenActive data. This will allow data publishers to identify and target areas for improvement and enrichment.

The use case focussed approach allows for more meaningful discussions and decisions about the quality of data, rather than simply labelling data as “good” or “poor”. It can also help us test the suitability of OpenActive to support new use cases, demonstrated in the diagram below. It shows how DQ metrics can indicate the data may be suited to one purpose, but less so to another.



Initial Focus

The initial focus of this data quality reporting framework is on the core use cases of the initiative:

Use Case 1: Discovery
In order to find a suitable activity nearby, a user needs to know several things.

What - indicated by the name, description and activity fields
Where - indicated by either the lat/long coordinates or the location’s postcode
When - indicated by the start and end dates and times
How much - indicated by either the price (offer) or the flag to indicate zero cost

Additional information required to decide if an activity is appropriate might include the level of difficulty (e.g. beginner, advanced) and any restrictions (e.g. Female only, Adults only), etc. These items will be explored further and community views on additional fields to include are welcomed.

Use Case 2: Booking
In addition to the above, a critical data element for an effective booking experience is a URL that takes the user directly to a booking page for the chosen activity. This may not be available for all activities. In many cases the booking process can be more complicated for example where the price of an activity may depend on whether a user is already a registered member with the activity provider.

Defining ‘good’ data quality

As a community of data publishers and data consumers, we have considered an ideal state for the data shared to support each use case. This state can be expressed in business rules and we can develop data quality tests to compare the data against the rules.

There are a number of data quality ‘dimensions’ that are helpful in describing the ideal state and developing appropriate measures:

Completeness: Are all the critical data elements present? Are all the relevant records included in the dataset? What is the impact of missing data? In OpenActive data, we have identified several fields to test for completeness, including the activity type, the location, and cost.

Consistency: Are values consistent within and between datasets? What are the impacts of inconsistent data? In OpenActive data, it is important that activities in the data feeds are consistent with the Activity List.

Currency/timeliness: Are records in the data feeds kept up to date in a timely manner? In OpenActive data, it is particularly important to update the availability of sessions or facilities to enable a smooth booking experience. Some feeds contain historic data, which must be filtered before you can display current opportunities in an activity finder.

Uniqueness: Identifiers play an important role in linking and processing OpenActive data. What is the impact of duplicated information? Unique URL links could be a proxy measure for a dedicated booking page (where a URL template is not already specified.

Validity: Are datasets, fields and values in the expected formats or ranges? The OpenActive validator tools perform many checks of validity but may be other tests we can explore e.g. a postcode may be present and may be in the correct alphanumeric format, but it may not be a ‘valid’ postcode.

It is important to recognise that these DQ metrics can not always capture the reality of an ideal state or reflect the systems and conditions that data publishers are working with. They are intended to broadly describe the current state of data in the OpenActive ecosystem to identify areas for potential improvement in relation to the specified use cases.

An example data quality dashboard with some proposed measures below as the basis for further discussion.

Example Data Quality Dashboard

Data Feeds	Sessions / Courses	Snapshot Date
222	373,00	Mid November 2022

Sessions / Courses

Activity Label matches the OA Activity List	Session or Course has a Description	Session or Course has an Activity Label, Name or Description
70%	60%	80%
Knowing where an activity fits helps activity finder developers create better interfaces, e.g. grouping activities to simplify search and navigation.	Description is a free text field which allows an activity provider to provide details to engage participants and help them decide if the activity is right for them.	Ideally a record should include one of these three to provide the 'what' of the opportunity.

Session or Course has geographic coordinates or a valid postcode	Session or Course has a future start date	Session or Course has a unique booking URL
70%	60%	42%
Activity finder developers can use postcodes or coordinates to display opportunities on a map and to search by location.	Activity finders focus on upcoming events.	Having a URL link directly to a booking page for the specific session significantly improves the user experience, though this is not possible in every case.

Session or Course has a cost or indication that it is free to attend.	Session or Course includes any relevant restrictions (e.g. age or gender)
X%	X%

Notes:

- The figures here are indicative estimates based on an initial analysis of data collected from 222 live OpenActive feeds in late November 2022. The calculations will be refined and quality assured in collaboration with data publishers.
- Metrics describing individual data feeds will allow publishers to identify potential issues and more effectively target improvements over time.
- These figures shown here are drawn from analysis of Scheduled Sessions and Courses. They can be expanded to cover other 'Event' types. Similar metrics will be developed for facilities (e.g. squash courts, multi-sport spaces).
- The records are not filtered by date - so some historic sessions will be included.

Next Steps

Sharing this framework is the first step in a longer process to improve the data quality across the OpenActive initiative. The data quality work will accompany other work streams including a review of the existing standards and the development of a use case framework.

This version of the data quality reporting framework is shared and open for comments from across the community - data publishers, data intermediaries, data consumers, advocates and end users.

Comments, ideas and options are specifically welcomed on the proposed set of DQ measures and the best means to publish these. Currently, the preferred approach is to present data quality indicators by feed in an updated status page.

The measures and dashboard will be refined over the coming months in collaboration with the community, with a final version of the data quality reporting framework going live in June 2023.