# **ReadMe:**

***Drug Orphanisation – Dataset 2: YouTube rare disease issue-networks***

This dataset was created and deposited onto the University of Sheffield Online Research Data repository (ORDA) on 21-Oct-20 21 by Dr. Matthew S. Hanchard, Research Associate at the University of Sheffield iHuman institute. It forms part of the work undertaken for the Wellcome Trust funded project ‘Orphan drugs: High prices, access to medicines and the transformation of biopharmaceutical innovation’ [219875/Z/19/Z]. This dataset comprise of two .csv format files used within workstream 2 of the project. They data appear in various outputs, e.g. publications and presentations. The deposited data were gathered using the University of Amsterdam Digital Methods Initiative’s ‘Data Tools for YouTube’ tool (DTFY) – which is available at <https://tools.digitalmethods.net/netvizz/youtube/> - before being processed and extracted from Gephi (0.9.2) – available at: <https://gephi.org/>. ORDA has full permission to store this dataset and to make it open access for public re-use without restriction under a CC BY license, in line with the Wellcome Trust commitment to making all research data Open Access. The University of Sheffield are the designated data controller for this dataset.

DTFY queries YouTube’s Application Programming Interface (API) using SQL and retrieves data on a pre-set text query, making it available as a .gdf format file. The query uses the search/list#relatedToVideoId endpoint of YouTube’s API, returning videos on a set search term and any videos related to it, i.e. those watched immediately after one with the search term in their title. Opening the DTFY .gdf file in Gephi allows for .csv files to be extracted on nodes an edges separately (as deposited here). This dataset used DTFY’s video network query on the search term ‘rare disease’ (with no parameters set to filter or limit results by date or geolocation).

Overall, the dataset consists of:

(1) a list of 7,469 individual videos with unique URLs posted to YouTube between 28-Jun-2006 and 19-May-2021. This file as titled ‘Orphan Drugs - Dataset 2 - YouTube rare disease issue-network – Nodes.csv’. The dataset includes a column called ‘Anonymity Level’ and another called ‘Label’. Where a channels/subscriber was an institution or public figure, their original name has been retained. Where consent to be named was granted, their original name has been retained. In all other instances they have been provided with a pseudonym (as is notable in the ‘Anonymity Level’ column).

(2) a list of 72,927 edges as links between the nodes (containing no identifiable information) , i.e. where one video has been watched after another. This file is titled ‘Orphan Drugs - Dataset 2 - YouTube rare disease issue-network – Edges,csv’.

The above process of data gathering, processing, and deposition aligns fully with YouTube’s service user terms and conditions. It also aligns with the ethical approval granted by the University of Sheffield School of Sociological Studies Research Ethics Committee (040659) on 14-Jun-2021.