Organising and validating data

Course Overview

Summary

Data can be used to create value; this value is recognised by stakeholders when we communicate the key insights and findings in the data. The value of data increases as it turns into a story. This is why data science and journalism skills must be used cooperatively. Those who work with the data must be able to communicate why it is important or useful, otherwise the value is lost. To ensure we maximise that value, we need to ensure that data is properly gathered, organised, structured and cleaned.

Creating value from data is a sequence. This sequence repeats in a cycle and follows the order: Gathering, Organising, Structuring and Cleaning.

Learning Outcomes

By the end of this course you will understand how to organise and clean data, as well as undertake activities to quality check and enable greater utility and interoperability of data.

To achieve this, you will:

- Investigate the gathering of data, including various sources of data and the practices employed in finding the data
- Examine the structure of data and determine approaches to organise the data into the appropriate structure as well as apply cleaning techniques to increase the utility of data
- Understand the purpose of data schemas and how to create and use them

Learning Experience

Number of modules	6 (+ reflective workbook + optional module)
Modality	Asynchronous / Self-directed / Online
Notional learning hours	3.5 hours (total)
Assessment	Formative
Certificate	Certificate of completion

Each module contains learning content that introduces the key concepts in the module, providing examples and case studies that demonstrate these concepts in practice. Each module contains a series of formative questions to support your learning. Learning is applied in activities throughout, and supports learners in cleaning data in order to improve its quality, accuracy and reliability, gain confidence in working with and managing data, and apply skills through a data structuring activity with feedback delivered by the Assistant AI Tutor.

Module Summary

Module Name	Description
Gathering data	Data plays a key role in storytelling, however not all data is easy to find. The growing demand for data has led to an increase in human friendly data services, including data portals and simple download buttons. However, downloadable data represents only a small fraction of the available data on the Web. The majority of data available on the Web is hidden from the human eye. But machines can find and read this data. In this section we explore the following: • The difference between downloadable and hidden data • Finding downloadable data • Finding hidden data • The benefits of using hidden data
Organising and structuring data	Once we have gathered data, we need to ensure it is usable. A common challenge in ensuring data quality is difficulties people face when using spreadsheets. When data is properly managed, it is much easier to answer fundamental questions and conduct analysis. Correctly managed data also ensures you aren't making decisions based on faulty evidence. Knowing how to structure and organise data in a spreadsheet is fundamental to ensuring consistency in your data. In this section we look at how to effectively structure a spreadsheet for raw data collection. We will cover: • Spreadsheet layouts • Column titles • Header rows and the freeze function • Data types
Choosing the right structure and format for data	How should data be structured? Who needs to consume the data - humans, machines or both? So, how should data be structured? Choosing the right format helps ensure the data can be simply managed and reused. To maximise reuse of data, it

	may be necessary to use a number of structures and formats available across different platforms to suit users' needs. In this section we'll explore: • The difference between machine and human-readable data • How to create machine-readable tabular data • Other data structures and formats • How to choose the right structure and format
Cleaning data	One of the biggest challenges when working with any data is dealing with errors. Often errors are not even noticed by data publishers because the data can change over many years. In other cases, errors can be the result of human mistakes in data entry, like mistyping or incorrect abbreviations. When working with any data, it is important to know how to find errors and correct them to make the data more useful. In this section we'll explore the following: • Common data errors • Useful data cleaning tools • Reasons for cleaning data
[Optional] Getting started with OpenRefine	Whenever you want to carry out analysis, you should always take the time to inspect the data you have been given to see if there are any errors like misspelling, repeated entries, mixed numerical scales and mixed ranges, as well as data that might be missing. Otherwise you run the risk of transferring these mistakes into your analysis. Open Refine is a powerful desktop app used for re-organising and cleaning data.
Hands-on: Organising and structuring data **(Assistant AI Tutor)**	Throughout this module you have been building your skills in the gathering, organising, structuring and cleaning of data. These skills are transferable across all different types of data.

	The following assignment has been designed to allow you to apply your skills in organising, structuring and cleaning a dataset. The assignment requires you to: • Reorganise and restructure a messy dataset into a single dataset that can be analysed simply by both human and machine. • Create a single CSV file (you can export this from tools like Excel) that contains just the dataset. • The dataset should contain enough data to enable analysis of the data by country, type, species and ownership.
Choosing and designing schemas	Data schemas are structured definitions or models that specify the structure, format, and constraints of data within a particular system or database. They define the organisation and expected characteristics of the data, including the data types, relationships, and rules that govern the data. Data schemas provide a blueprint for how data should be structured and stored, ensuring consistency and integrity. While they provide similar features to data validation in Excel, data validation rules in Excel are primarily applied to individual cells or ranges and focus on immediate data validation within the spreadsheet. However, they do not provide a comprehensive and standardised structure for an entire dataset or database. In this section we explore: • What is a schema? • Why do schemas matter? • Schemas in practice • Choosing or designing schemas