

DATA SOURCES

ACCESSING DIFFERENT SOURCES OF DATA

OUTLINE

- Course Outline
- Basic concepts
 - Data sources and accessing different sources of data
 - Data Resolution
- Metadata
 - Concepts
 - Standards
 - Paradata
- Data Provenance
 - Understanding provenance for the IMD
- Data Quality

DATA SOURCES

ACCESSING DIFFERENT SOURCES OF DATA

Most common data types [1]

- Data can be classified according to its main production source:
- Primary data when data is generated methods developed and implemented by the user
- Secondary data when data is collected from databases which were processed and made available by third parties

Primary Data [1]

- Primary data is data generated by the research
- Multiple methods to generate primary data
 - Surveys
 - Interviews
 - Observations
 - Data mining methods
 - ...

Primary Data [2]

- Control over the data collected
- Raw data that needs to be cleaned, structured and analysed
- Allows the application of data analysis methods tailored to the research
- Implies a full consideration of ethical and legal issues around the data generated

Secondary Data [1]

- Secondary data is data that has been processed by a previous agent
- Can be qualitative or quantitative
- Has a given structure
- Attributes were previously selected
- May or may not have issues of privacy
- Should have been cleaned of errors, inconsistencies or missing values

Secondary Data [2]

- Datasets usually used in various remits of public policy
- Validated and cleaned
- Usually free from ethical considerations around privacy, needs to be referenced
- Useful in research, comparability
- Useful for triangulation

Resolution [1]

■ Temporal resolution

- Different datasets tend to have different temporal resolutions, posing challenges to create common datasets

■ Spatial resolution

- Different datasets usually have different resolutions
- Critical issue in quantitative spatial data
- Data carries evidences of phenomena that can be identified at a given scale
- Aggregations of data (scaling up) or disaggregation (scaling down) may lead to lost of representativeness
- The Modifiable Area Unit Problem (MAUP) and the Ecological Fallacy

Resolution [2]

- Temporal resolution

- 5, 10 years – population censuses, employment, deprivation
- Day, month year – individual data, business data
- Millisecond, second, hour – market data, business data, transport/communications data

- Spatial resolution

- Individual
- Business
- Trips
- Geographical units – output areas, LSOA, MSOA, Local Authorities, ...

Modifiable Area Unit Problem [1]

MAUP is a source of bias that can radically affect data analysis and interpretation

MAUP occurs when data are aggregated up (often from individual points) and the resulting summary values (e.g. totals, rates, proportions) are influenced by the choice of boundaries to which the data is aggregated.

It is the variation in the spatial units used for aggregation that causes variation in statistical results.

The choice of boundaries has implications for classifying areas

[See Openshaw, 1981]



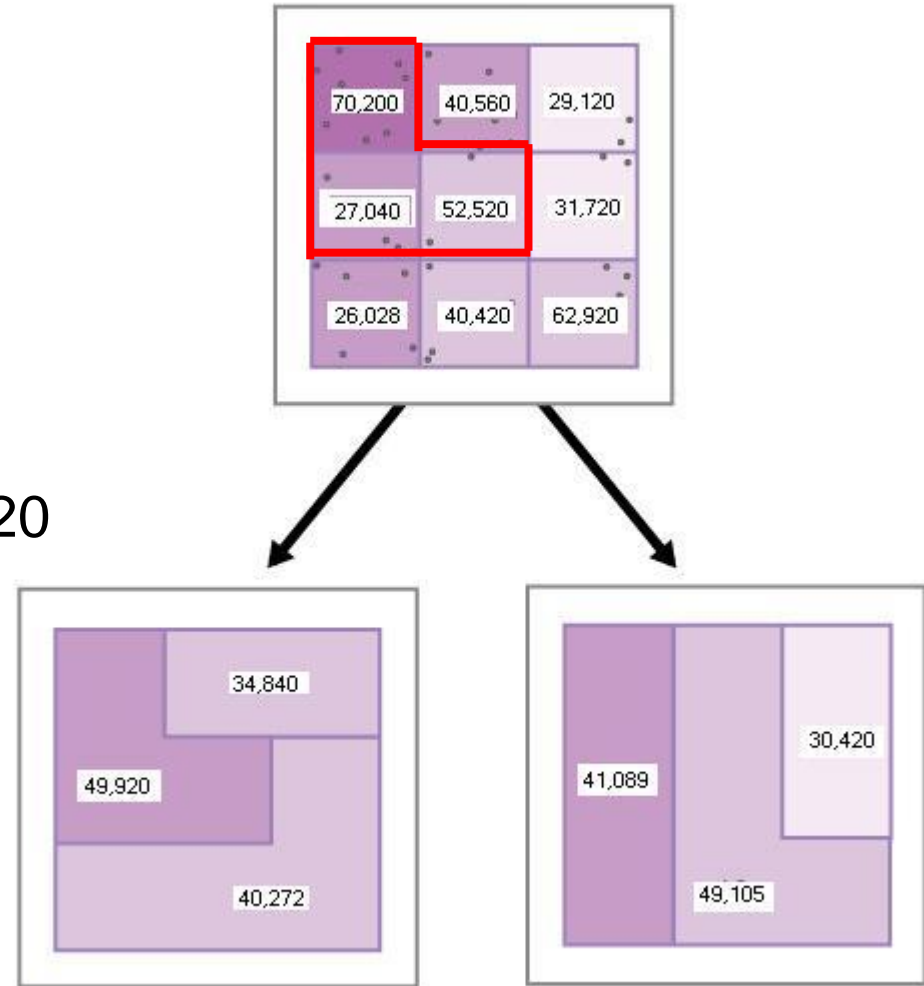
Modifiable Area Unit Problem [2]

Blocks represent a census unit

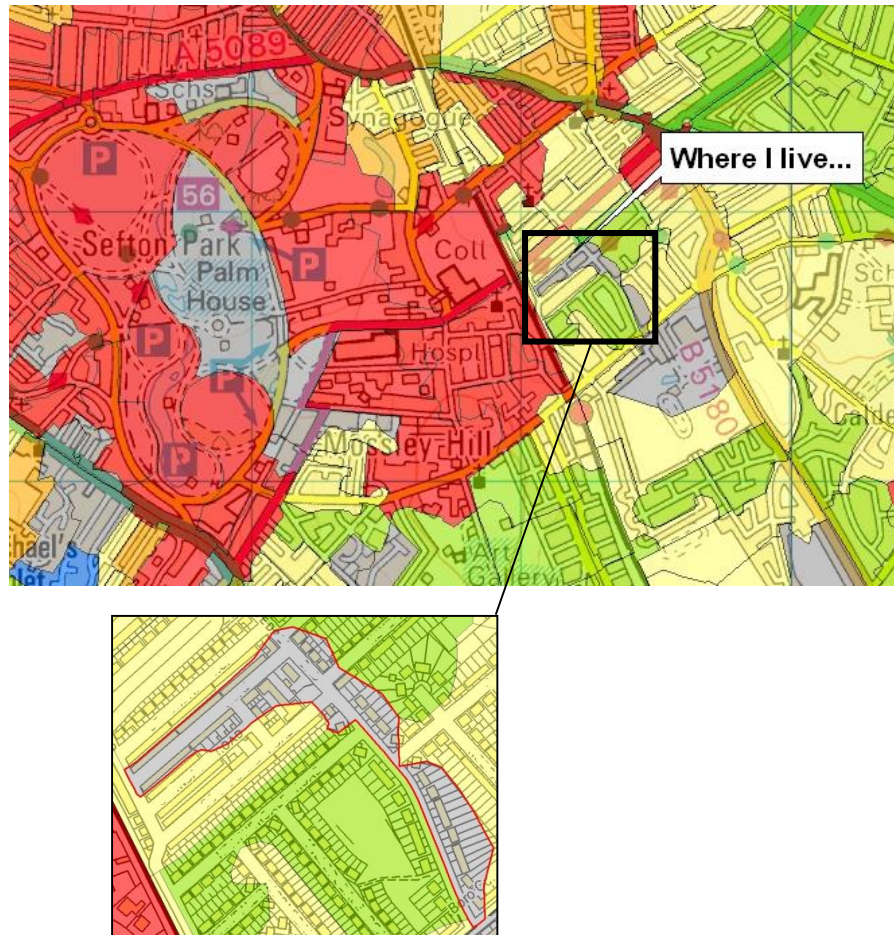
Data: Model-based estimates of average household income for each census unit

$$\frac{70,200 + 27,040 + 52,520}{3 \text{ (no of blocks)}} = 49,920$$

Significantly less than 70,200
and significantly greater than
27,040!



Ecological Fallacy



Ecological Fallacy: “The mistake of assuming that where relationships are found among aggregate data, these relationships will also be found among individuals or households”.

Group 5 – *Constrained by circumstances*

Group 5b – *Older workers*

#‘1’ – Stephen (30-something lecturer)

#‘3’ – ‘James and Louise’ (‘hipsters’ early 20s)

#‘15’ – ‘Jack and Mary’ (wealthy pensioners)

[See Voas and Williamson, 2001; Harris, 2001; Batey and Brown, 2007]

Change of Support Problem

COSP refers to issues of compatibility of different variables measured at different scales

A hierarchy of geographies might exist but usually they don't match, example ONS Lower Super Output Areas from 2001 and 2011

To tackle COSP need to find a common scale of aggregation or use interpolation/imputation methods

METADATA AND PARADATA

LEARNING ABOUT THE DATASETS

Definitions

- Metadata is broadly speaking “data about data”
- “meta” is the Greek prefix for beyond
- so Metadata is Data that provides information about a specific data set, focusing on different aspects such as source, resolution, production or accuracy
- Has multiple purposes
 - To manage the datasets
 - To report on information
- It's stored in metadata repositories, which are databases in itself

Metadata General Properties

- **Sufficiency**
 - Can an object describe itself?
 - e.g., images
- **Scalability**
 - Allows for rapid searching
 - Searching metadata fields vs large data files
- **Interoperability**
 - Can exchange data using mutually agreed metadata formats

Types of metadata

- Types of metadata
- Structural or guide metadata Bretheron and Singley (1994)
 - Structural metadata describes the datasets (tables, attributes, indexes, etc)
 - Guide metadata describes contents in common language
- Technical or business metadata Kimball (1996)
 - Technical metadata for the main internal attributes
 - Business metadata for the external attributes

Use of Metadata

- **Users** (or potential users):
 - search across the range of resource descriptions made available by different resource providers, regardless of the fact that those descriptions may use the conventions of different resource description communities.
 - combine or compare descriptions of resources from different communities
 - reuse both the resources and the descriptions of those resources in new contexts
- **Providers:**
 - disseminate descriptions of their resources to potential users and often as widely as possible
 - share descriptions of their resources with other resource description communities
 - describe relationships between their resources and those of other resource description communities
- **Third parties:**
 - describe (or add to existing descriptions of) resources owned by others
 - describe relationships between resources from multiple resource description communities

Common Standards for Metadata

- Dublin Core
- ISO 19115 for geographic information
- Data Documentation Initiative (DDI)
- Statistical Data and Metadata eXchange (SDMX)
- Metadata Encoding and Transmission Standard (METS)
- General International Standard Archival Description (ISAD(G))
- DataCite metadata schema for the publication and citation of digital datasets with a persistent identifier.

Basic data attributes about the data

- Title
- Date
- Subject descriptors
- Creator(s) (Creator of the dataset; main researchers involved)
- Funders
- File format
- Storage location of the data (including identifier information)
- Origin of the data (creation/acquisition of the data)
- Time references for the data (key dates associated with the data: start, end, release, etc)
- Access conditions
- Terms of use of the data

Metadata Quality (next week)

- Missing metadata
- Ambiguity
- Unreadability
- Inaccuracy

Discussion (15 minutes)

Critically think about the importance of metadata as a management tool for dealing with data. Think about your own research interests and search in the Research Data Alliance the standards and tools available for your area of interest. Access the list of subject areas at <http://rd-alliance.github.io/metadata-directory/subjects/>

Accessing and managing metadata

The screenshot shows a web browser displaying the rd-alliance.github.io/metadata-directory/tools/ page. The page has a sidebar on the left with the title "Metadata" and the subtitle "RDA | Metadata Directory". The sidebar contains links for "Edit this page", "View the standards", "View the extensions", "View the tools", "View the use cases", "Browse by subject areas", "Contribute", "Add standards", "Add extensions", "Add tools", and "Add use cases". The main content area is titled "Social and Behavioral Sciences" in red. It lists several tools, each with a link to the tool and an "Edit" button. The tools listed are: "DDI Tools", "DDI on Rails", "FISH Interoperability Toolkit", "Istat SDMX Framework Project", "SDMX Editor", "SDMX Mapping Assistant", "SDMX Tool Repository", and "Converis". Each tool description is provided below its name. The browser's address bar and bookmarks are visible at the top.

Metadata
RDA | Metadata Directory

Edit this page

View the standards
View the extensions
View the tools
View the use cases
Browse by subject areas

Contribute
Add standards
Add extensions
Add tools
Add use cases

Social and Behavioral Sciences

[DDI Tools](#) [Edit](#)

The Data Documentation Initiative website's list of tools to implement the [DDI](#) standard.

[DDI on Rails](#) [Edit](#)

Server-side software for building a data portal, with a particular focus on survey datasets. It uses DDI to provide access to the data at the level of concepts and variables. For an example of it in use, see the [SOEPinfo data portal](#).

[FISH Interoperability Toolkit](#) [Edit](#)

A suite of tools using the [MIDAS](#) Heritage metadata standard to facilitate the process of moving information between the wide variety of information systems used to record the historic environment.

[Istat SDMX Framework Project](#) [Edit](#)

A suite of tools for managing data and metadata in [SDMX](#).

[SDMX Editor](#) [Edit](#)

A simple tool for managing and accessing statistical metadata, using the [SDMX](#) framework.

[SDMX Mapping Assistant](#) [Edit](#)

A tool to facilitate the mapping between the structural metadata provided by an [SDMX-ML](#) Data Structure Definition and those that reside in a database of a dissemination environment.

[SDMX Tool Repository](#) [Edit](#)

A list of software tools supporting the [SDMX](#) standard.

General Research Data

[Converis](#) [Edit](#)

Current research information system implementing the CERIF standard. Originally developed by Avedas but now a product of Thomson Reuters.

EU INSPIRE Initiative

Data Specifications > Themes

Secure | <https://inspire.ec.europa.eu/Themes/Data-Specifications/2892>

Apps | Add to My Bookmarks | The Knowledge | Foot | Eltis | The urban mobil | GeoDa Software - YouT | Window seat or aisle - | Hemidido | * Calendar | Trello | Google Street View - E | Other bookmarks

About cookies

This site uses cookies to offer you a better browsing experience. Find out more on [how we use cookies](#) and [how you can change your settings](#).

Accept

About | Contact | Terms of use | Privacy Policy | Legal Notice | Cookies

English (en)

INSPIRE KNOWLEDGE BASE

Infrastructure for spatial information in Europe

European Commission > INSPIRE > Implement > Data Specifications > Themes

Home | Learn | **Implement** | Participate | Use | Toolkit

Implement









- Guide for implementers
- Roadmap
- Data Specifications**
- Monitoring & Reporting
- Metaddata
- Network Services
- Data and Service Sharing
- Spatial Data Services
- INSPIRE Coordination
- Maintenance and Implementation Framework

Data Specifications

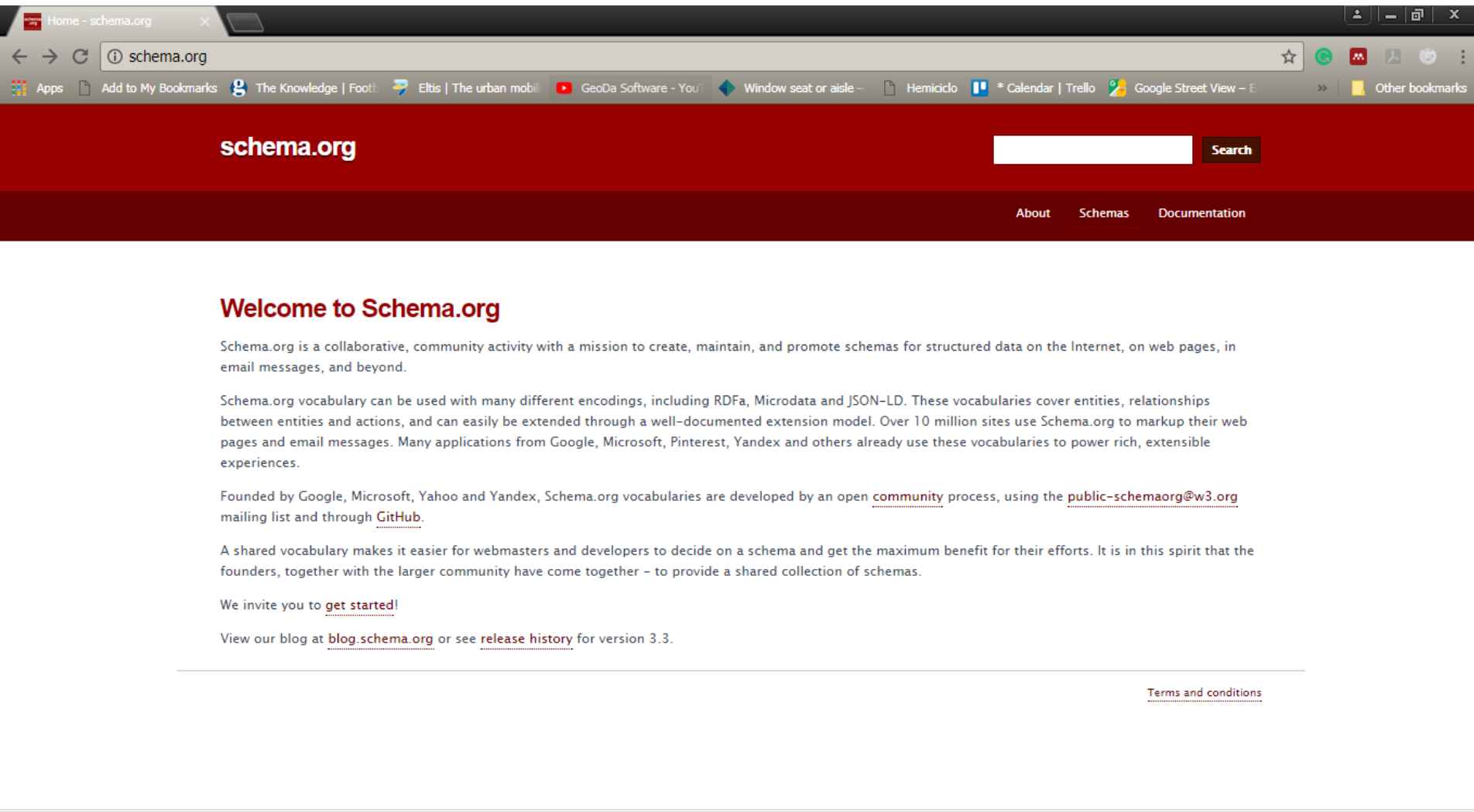
Overview

Data Specifications > Themes

ANNEX: 1

-  [Addresses](#)
-  [Cadastral parcels](#)
-  [Geographical grid systems](#)
-  [Hydrography](#)
-  [Administrative units](#)
-  [Coordinate reference systems](#)
-  [Geographical names](#)
-  [Protected sites](#)

Schema.org



The image is a screenshot of a web browser displaying the Schema.org homepage. The browser's address bar shows 'schema.org'. The page has a dark red header with the 'schema.org' logo on the left and a search bar on the right. Below the header, there are navigation links for 'About', 'Schemas', and 'Documentation'. The main content area is white and features a 'Welcome to Schema.org' section. This section includes a paragraph about Schema.org's mission, a paragraph about its use with various encodings, a paragraph about its founding by Google, Microsoft, Yahoo, and Yandex, and a paragraph about its shared vocabulary. At the bottom of the main content area, there are links to 'get started!', 'blog.schema.org', and 'release history'. A footer link for 'Terms and conditions' is located at the bottom right of the page.

Home - schema.org

schema.org

Apps Add to My Bookmarks The Knowledge | Footl. Eltis | The urban mobil. GeoDa Software - You Window seat or aisle - Hemiciclo * Calendar | Trello Google Street View - E Other bookmarks

schema.org

Search

About Schemas Documentation

Welcome to Schema.org

Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.

Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD. These vocabularies cover entities, relationships between entities and actions, and can easily be extended through a well-documented extension model. Over 10 million sites use Schema.org to markup their web pages and email messages. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.

Founded by Google, Microsoft, Yahoo and Yandex, Schema.org vocabularies are developed by an open [community](#) process, using the [public-schemaorg@w3.org](#) mailing list and through [GitHub](#).

A shared vocabulary makes it easier for webmasters and developers to decide on a schema and get the maximum benefit for their efforts. It is in this spirit that the founders, together with the larger community have come together - to provide a shared collection of schemas.

We invite you to [get started!](#)

View our blog at [blog.schema.org](#) or see [release history](#) for version 3.3.

[Terms and conditions](#)

Data Definition [1]

The screenshot shows a web browser window with the URL `infuse2011.ukdataservice.ac.uk/InFuseWiz.aspx?cookie=openaccess`. The browser's address bar and tabs are visible at the top. Below the browser window, the website's interface is shown. It features a 'Steps' navigation bar with four steps, where step 1 is highlighted. The main heading is 'Topics'. A note below the heading states: 'Some topics are only available for certain countries; E - England, N - Northern Ireland, S - Scotland, W - Wales'. The content area is titled 'Showing selected topic combination for topic(s) (Age, National identity)'. It contains three sections: 'Age', 'National identity', and 'Sex'. The 'Age' section explains that age is derived from the date of birth question and is a person's age at their last birthday, as of 27 March 2011. It notes that dates of birth implying an age over 115 are treated as invalid and the person's age is imputed, and that infants less than one year old are classified as 0 years of age. The 'National identity' section explains that a person's national identity is a self-determined assessment of their own identity with respect to the country or countries with which they feel an affiliation. It notes that this assessment is not dependent on legal nationality or ethnic group. The 'Sex' section explains that the classification of a person is either male or female. The 'Unit' section explains that the unit is for a particular count (e.g. people or households). At the bottom of the content area, there is a 'Previous' button. The Windows taskbar is visible at the very bottom of the image, showing the Start button and several application icons.

Steps: 1 2 3 4

Topics

Some topics are only available for certain countries; E - England, N - Northern Ireland, S - Scotland, W - Wales

Showing selected topic combination for topic(s) (Age, National identity)

Age

Age is derived from the date of birth question and is a person's age at their last birthday, at 27 March 2011. Dates of birth that imply an age over 115 are treated as invalid and the person's age is imputed. Infants less than one year old are classified as 0 years of age.

National identity

A person's national identity is a self-determined assessment of their own identity with respect to the country or countries with which they feel an affiliation. This assessment of identity is not dependent on legal nationality or ethnic group. The national identity question included six tick box responses: one for each of the four parts of the UK (English, Welsh, Scottish, Northern Irish), one for British, and one for 'other'. Where a person ticked 'other' they were asked to write in the name of the country. People were asked to tick all options that they felt applied to them. This means that in results relating to national identity people may be classified with a single national identity or a combination of identities.

Sex

The classification of a person as either male or female.

Unit

The unit is for a particular count (e.g. people or households)

[Previous](#)

Data Definition [2]

Meta_AGE_NATIDE_SEX_UNIT - Microsoft Excel

A1	CDU_FIELD_NAME				
A	B	C	D		
1	CDU_FIELD_NAME	CENSUS_CELLNAME	TOPIC	CATEGORY	TOPIC_DESCRIPTION
2	F17285	DC2102EW0151	Age	Age 30 to 34	Age is derived from the date of birth question and is a person's age at their last birthday, at 27 March 2011. Dates of birth that imply
3	F17285	DC2102EW0151	National identity	Total\ National identity	A person's national identity is a self-determined assessment of their own identity with respect to the country or countries with whi
4	F17285	DC2102EW0151	Sex	Total\ Sex	The classification of a person as either male or female.
5	F17285	DC2102EW0151	Unit	Persons	The unit is for a particular count (e.g. people or households)
6	F17286	DC2102EW0481	Age	Age 30 to 34	Age is derived from the date of birth question and is a person's age at their last birthday, at 27 March 2011. Dates of birth that imply
7	F17286	DC2102EW0481	National identity	Total\ National identity	A person's national identity is a self-determined assessment of their own identity with respect to the country or countries with whi
8	F17286	DC2102EW0481	Sex	Males	The classification of a person as either male or female.
9	F17286	DC2102EW0481	Unit	Persons	The unit is for a particular count (e.g. people or households)
10	F17287	DC2102EW0811	Age	Age 30 to 34	Age is derived from the date of birth question and is a person's age at their last birthday, at 27 March 2011. Dates of birth that imply
11	F17287	DC2102EW0811	National identity	Total\ National identity	A person's national identity is a self-determined assessment of their own identity with respect to the country or countries with whi
12	F17287	DC2102EW0811	Sex	Females	The classification of a person as either male or female.
13	F17287	DC2102EW0811	Unit	Persons	The unit is for a particular count (e.g. people or households)
14	F18536	DC2102EW0181	Age	Age 40 to 44	Age is derived from the date of birth question and is a person's age at their last birthday, at 27 March 2011. Dates of birth that imply
15	F18536	DC2102EW0181	National identity	Total\ National identity	A person's national identity is a self-determined assessment of their own identity with respect to the country or countries with whi
16	F18536	DC2102EW0181	Sex	Total\ Sex	The classification of a person as either male or female.
17	F18536	DC2102EW0181	Unit	Persons	The unit is for a particular count (e.g. people or households)
18	F18537	DC2102EW0511	Age	Age 40 to 44	Age is derived from the date of birth question and is a person's age at their last birthday, at 27 March 2011. Dates of birth that imply
19	F18537	DC2102EW0511	National identity	Total\ National identity	A person's national identity is a self-determined assessment of their own identity with respect to the country or countries with whi
20	F18537	DC2102EW0511	Sex	Males	The classification of a person as either male or female.
21	F18537	DC2102EW0511	Unit	Persons	The unit is for a particular count (e.g. people or households)
22	F18538	DC2102EW0841	Age	Age 40 to 44	Age is derived from the date of birth question and is a person's age at their last birthday, at 27 March 2011. Dates of birth that imply
23	F18538	DC2102EW0841	National identity	Total\ National identity	A person's national identity is a self-determined assessment of their own identity with respect to the country or countries with whi
24	F18538	DC2102EW0841	Sex	Females	The classification of a person as either male or female.
25	F18538	DC2102EW0841	Unit	Persons	The unit is for a particular count (e.g. people or households)
26	F19787	DC2102EW0166	Age	Age 35 to 39	Age is derived from the date of birth question and is a person's age at their last birthday, at 27 March 2011. Dates of birth that imply

Meta_AGE_NATIDE_SEX_UNIT

Ready

Start

y_Apps » y_Shortcuts » EN

06:50

Terms and conditions of using data

TermsAndConditions.html

file:///C:/Users/mzdssnp2/Documents/2_Teaching/0_SEED_UGT_PGT/2017_2018_S2_PLAN70001_MSCDATA_UDIE/DATA/TermsAndConditions.html

This UK Data Service Census Support boundary dataset is made available to you under the terms of the [UK Open Government Licence](#) v3.

Your use of the UK Data Service is subject to the terms of the UK Data Service [End User Licence](#)

The following copyright statements MUST be used when reproducing or using this dataset, for example when displaying boundary data as a map or other graphic within a written report:

- Contains National Statistics data © Crown copyright and database right [year]**
- Contains NRS data © Crown copyright and database right [year]**
- Source: NISRA : Website: www.nisra.gov.uk**
- Contains OS data © Crown copyright [and database right] (year)**

Note, that where any of the above copyright statements contain the word year, this should be replaced with the year, for example 2015, in which you used the UK Data Service dataset or wrote your report etc.

In addition you can also include the following citation as a reference for this UK Data Service Census Support dataset where you have referred to this UK Data Service dataset in an essay, research paper or journal article:

Office for National Statistics;National Records of Scotland;Northern Ireland Statistics and Research Agency (2011). 2011 Census: boundary data (United Kingdom) [data collection]. UK Data Service. SN:5819 UKBORDERS: Digitised Boundary Data, 1840- and Postcode Directories, 1980-. <http://discover.ukdataservice.ac.uk/catalogue/?sn=5819&type=Data%20catalogue>, Retrieved from <http://census.ukdataservice.ac.uk/get-data/boundary-data.aspx>.

Contains public sector information licensed under the Open Government Licence v3.

If you have any questions about the use or licensing of this dataset please contact the UK Data Service via:

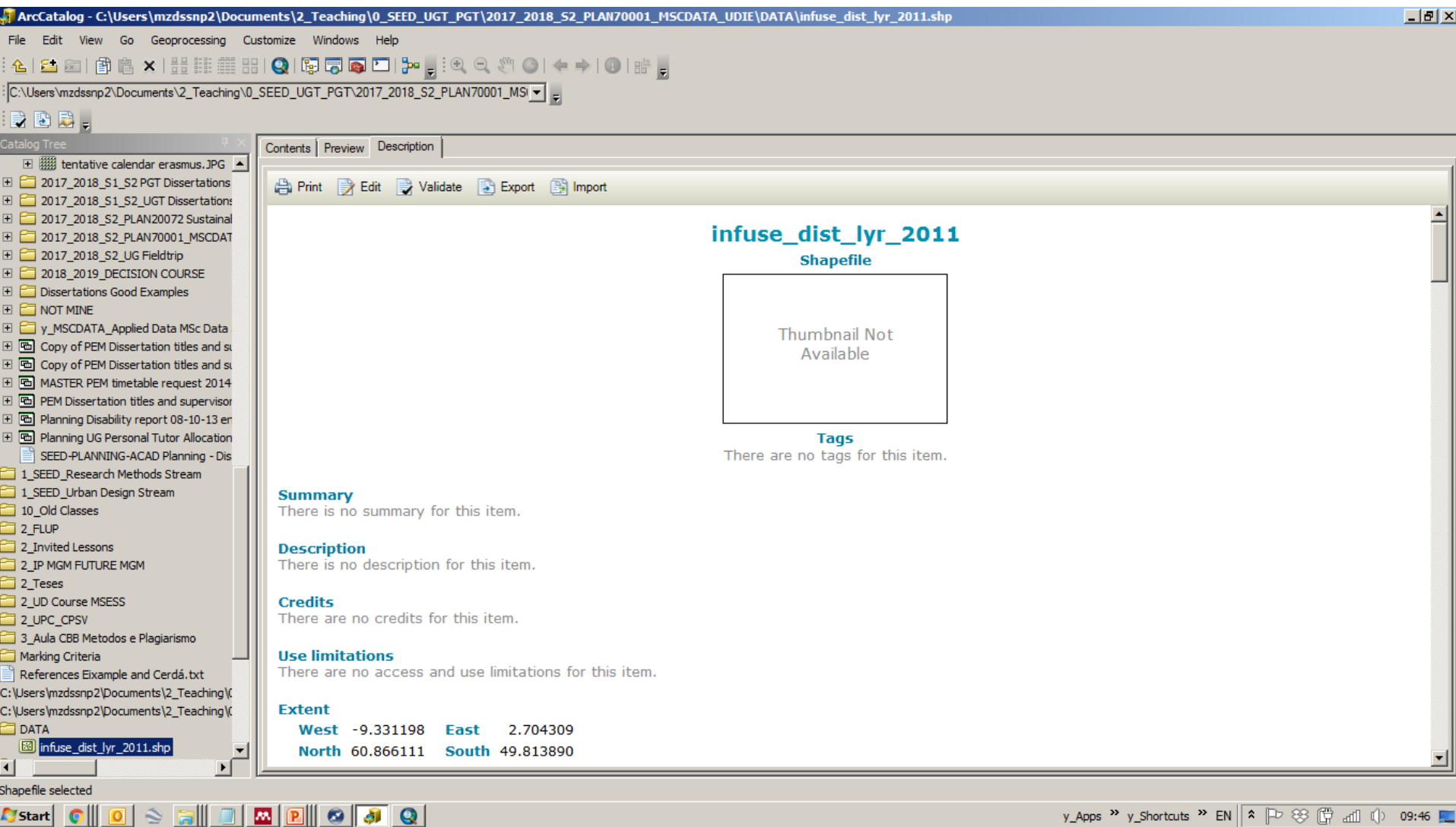
<http://ukdataservice.ac.uk/help/get-in-touch.aspx>

UK Data Service, November 2015

ukdataservice.ac.uk/help/get-in-touch.aspx

y_Apps » y_Shortcuts » EN 09:17

ARCGIS Metadata viewer



Paradata [1]

- “Paradata are auxiliary data about the process of data collection and include keystroke files and time stamps”

Couper M. New technologies and survey data collection: challenges and opportunities. 2002. Paper presented at: International Conference on Improving Surveys; Copenhagen, Denmark.

http://www.icis.dk/ICIS_papers/Keynote1_0_3.pdf

- “Paradata, which provide information about the survey data collection process (Couper, 2000a), lend insight into errors and costs that can impact the quality of a survey data collection—often at a low cost of collection to researchers.”

McCain et al (2019) A Typology of Web Survey Paradata for Assessing Total Survey Error, Social Science Computer Review Vol. 37(2) 196-213

Paradata [2]

- “Despite the problems currently associated with web surveys, such as coverage and nonresponse errors, web surveys have at least one particularly interesting feature: They record a wide array of paradata.”

Heerwegh, D. (2003). Explaining Response Latencies and Changing Answers Using Client-Side Paradata From a Web Survey. , 21(3), pp.360–373.

Paradata [2]

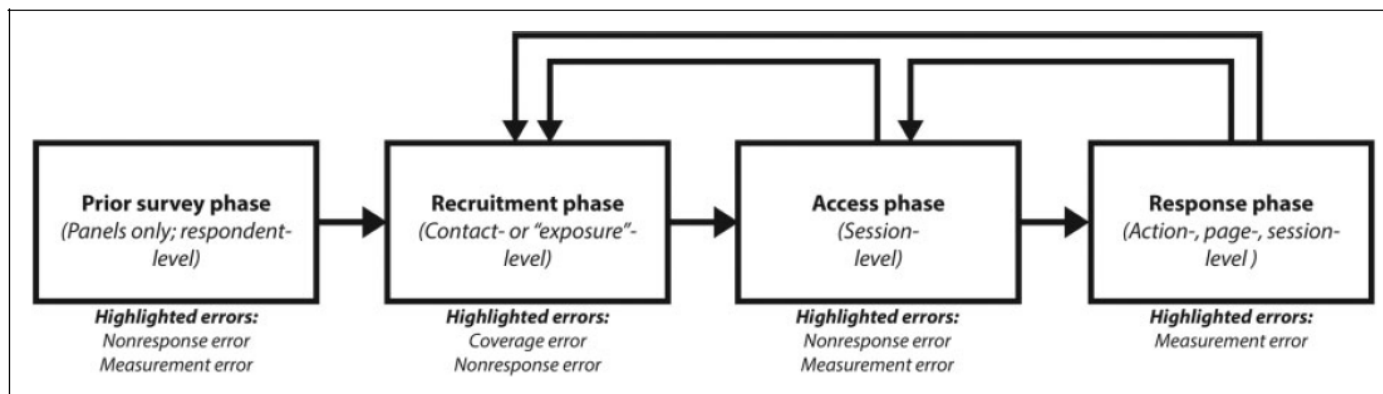


Figure 1. Phases of the web survey data collection process at which paradata can be collected.

Table 1. Examples of Paradata Available for Collection at Each Stage of the Data Collection Process.

Phase	Examples of Collected Paradata
Prior survey phase	Level-of-effort information (contacts) from previous waves Item missing data Response speed indices Previous mobile device use
Recruitment phase	Indicator of e-mail opening Times/dates of contacts Time to open contact Recruitment stream (river/intercept samples)
Access phase	Time from open contact to access Device characteristics (type, screen dimensions, resolution and orientation, etc.) Login attempts (total count, count of successes, etc.)
Response phase	Response times Item missing data Navigation (e.g., backups) and response changes Loss of focus/out of browser window Mouse movements, touch events, and keystrokes Device characteristics (type, screen dimensions, resolution and orientation, etc.)

Paradata [3]

- Server-side paradata: describes server events but not the respondent's action in the webpage
- Client-side paradata: collected at the respondent's client machine, its behaviour in the webpage, how it filled in the survey, though some scripting tool

Heerwegh, D. (2003). Explaining Response Latencies and Changing Answers Using Client-Side Paradata From a Web Survey. , 21(3), pp.360–373.

Paradata [4]

Paradata can be used to improve the management of data collection in the following ways:

- To achieve a better understanding of the data collection processes and to identify opportunities to improve survey operations;
- To evaluate new data collection initiatives;
- To monitor data collection and identify problems in a timely fashion (e.g. problems with the data collection instruments, problems with interviewer performance, possible interviewer falsification);
- To produce quality control metrics for internal use but also for sponsors;
- To produce more accurate information about costs and data quality;
- To inform survey design decisions about trade-offs between fieldwork costs, data quality and time;
- To make changes to the survey design during data collection in order to optimise the trade-off between costs, quality and time (i.e. responsive design)

Nicolaas, G. (2011). ESRC National Centre for Research Methods Review paper Survey Paradata : A review. , (January), pp.1–21.

Paradata [5]

- Classification of respondents

Sowan, A.K. (2010). A New Data Source From. Computers, Informatics, Nursing, 28(6), pp.333–342

Table 1

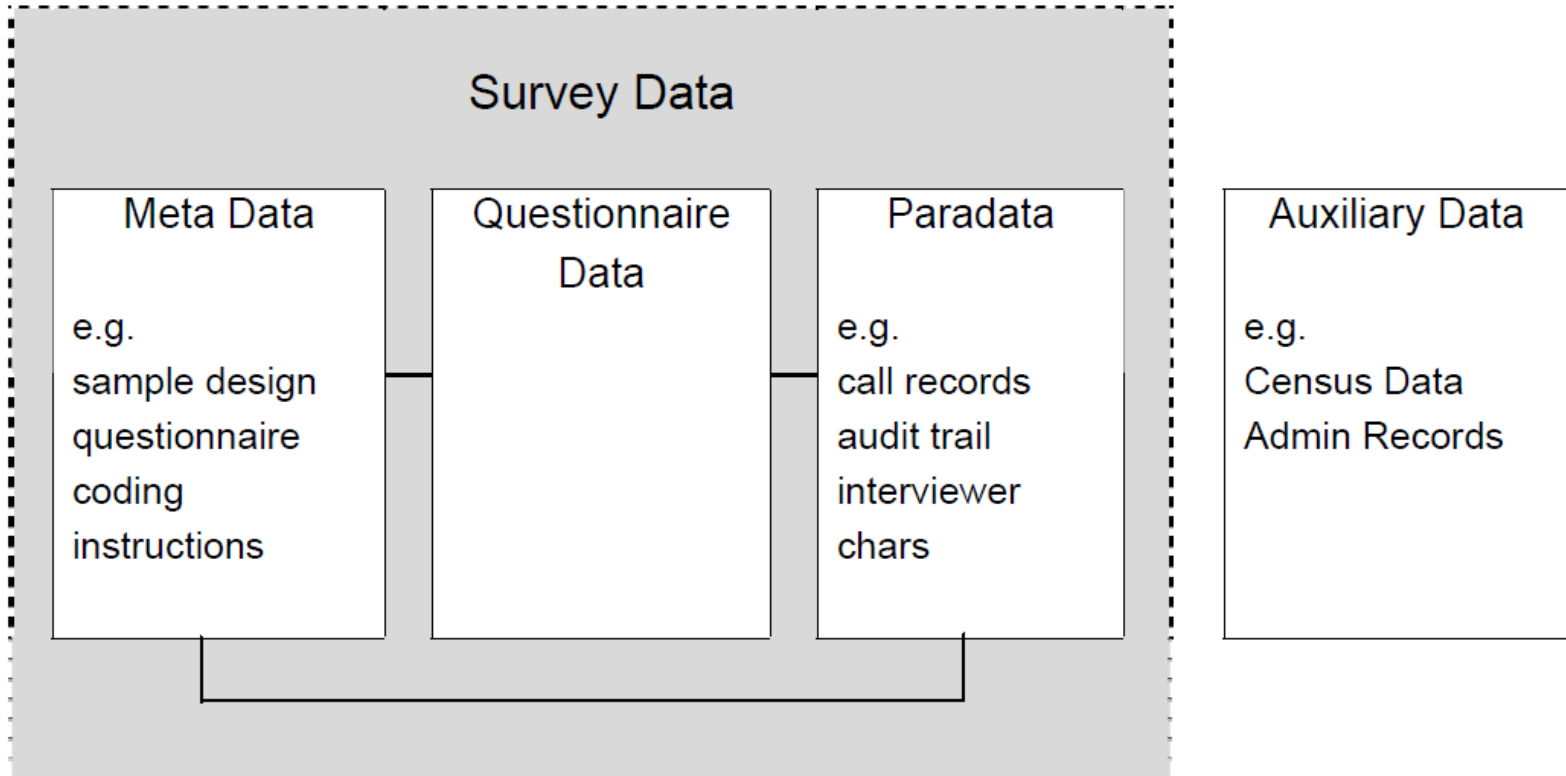
Response Patterns in Web-Administered Measures



Response Pattern	Definition
Complete responders	Subjects viewed and answered all items
Unit nonresponders	Subjects clicked the hyperlink of a measure and/or logged in but did not participate and did not view any item
Answering dropouts	Subjects dropped out after completion of some items but did not view all items
Lurkers	Subjects viewed all items but did not respond to any
Lurking dropouts	Subjects viewed some items but did not respond to any
Item nonresponders	Subjects viewed all items but responded to only some
Item nonresponders-dropouts	Subjects viewed some items and responded to only some of the viewed items

Paradata and Metadata

Figure 1.1 Diagram showing the three types of survey data and auxiliary data



Nicolaas, G. (2011). ESRC National Centre for Research Methods Review paper Survey Paradata : A review. , (January), pp.1–21.

DATA PROVENANCE

UNDERSTANDING HOW DATA IS ASSEMBLED

Concept

- “The term “data provenance” refers to a record trail that accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place.”

Gupta A. (2009) Data Provenance. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA

- Also referred to as lineage or pedigree

Example

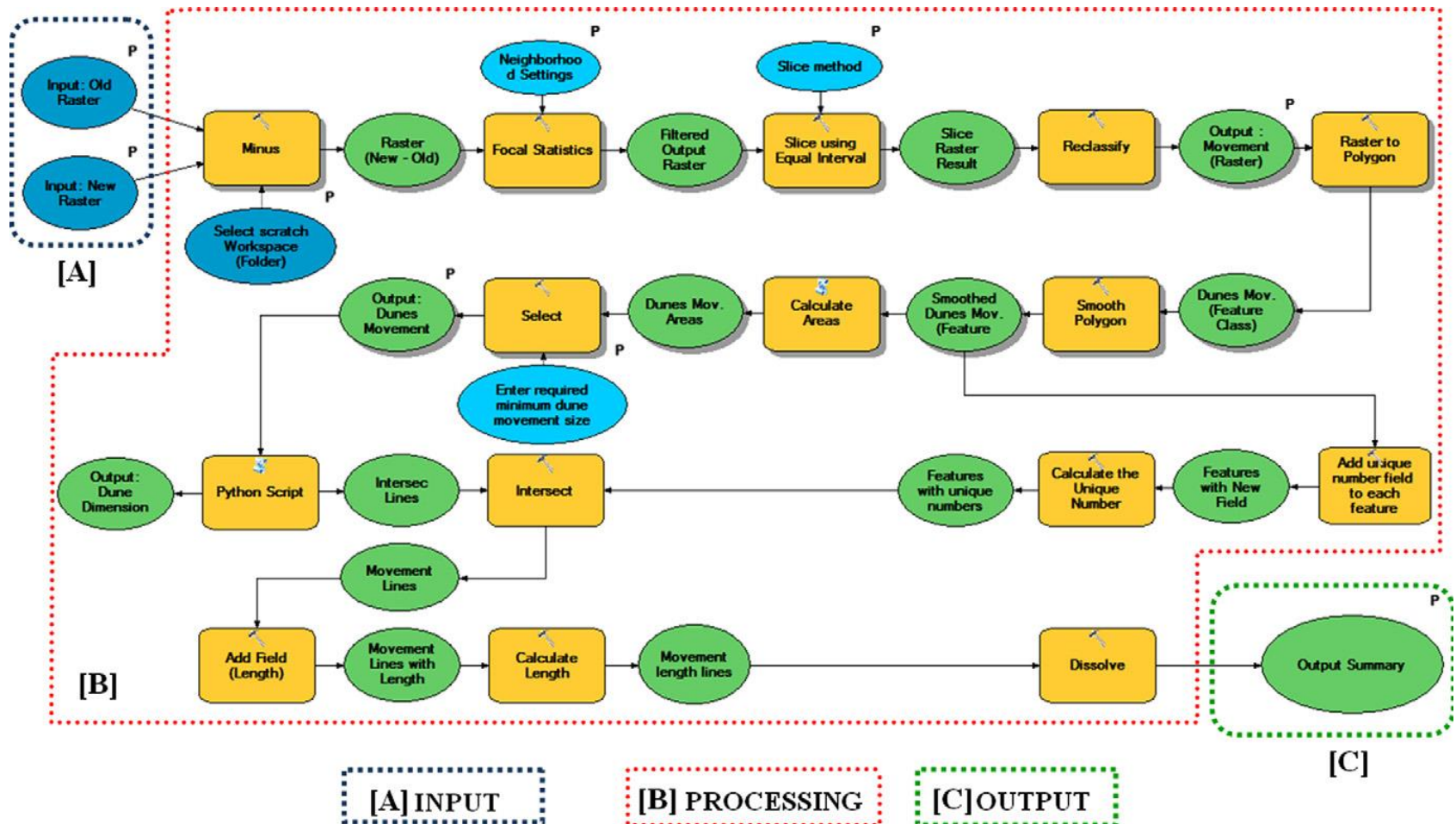
- “In an application like Molecular Biology, a lot of data is derived from public databases, which in turn might be derived from papers but after some transformations (only the most significant data were put in the public database), which are derived from experimental observations. A provenance record will keep this history for each piece of data”

Gupta A. (2009) Data Provenance. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA

Uses for data provenance

- Data Quality: estimate data quality and data reliability based on the source data and transformations
- Audit Trail: trace the audit trail of data, determine resource usage, and detect errors in data generation.
- Replication Recipes: Detailed provenance information can allow repetition of data derivation, help maintain its currency and be a recipe for replication
- Attribution: Pedigree can establish the copyright and ownership of data, enable its citation, and determine liability
- Informational: A generic use of lineage is to query based on lineage metadata for data discovery. It can also be browsed to provide a context to interpret data.

A flow model of data provenance

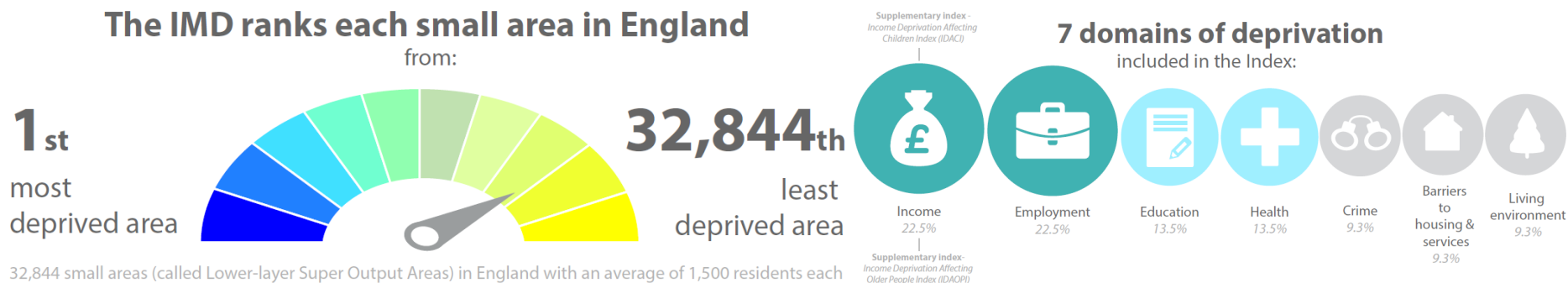


Ghadiry, M., Shalaby, A., Koch, B., 2012. A new GIS-based model for automated extraction of Sand Dune encroachment case study : Dakhla Oases , western desert of Egypt. Egypt. J. Remote Sens. Sp. Sci. 15, 53–65. <https://doi.org/10.1016/j.ejrs.2012.04.001>

Example IMD 2015 Crime Domain

Consider the Indices of Multiple Deprivation for 2015. Choose one domain and analyse the formation of the index considering its multiple sources. Use metadata when creating the data provenance flow.

All information about the IMD 2015 is available at <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015>



DATA QUALITY

ASSESSING THE USABILITY OF DATASETS

Definition

- Measuring data parameters against data standards to evaluate the level of quality of the data or dataset
- “Data quality is the capability of data to be used effectively, economically and rapidly to inform and evaluate decisions.” Karr and Sanil (2002)
- Includes multiple dimensions
 - A dimension is an data item, a record a dataset or a database that can be used as a parameter of data quality

Data quality dimensions

- Completeness
 - Validity
 - Accuracy
 - Consistency
 - Uniqueness
 - Timeliness
- 

Completeness [1]

Title	Completeness
Definition	The proportion of stored data against the potential of "100% complete"
Reference	Business rules which define what "100% complete" represents.
Measure	A measure of the absence of blank (null or empty string) values or the presence of non-blank values.
Scope	0-100% of critical data to be measured in any data item, record, data set or database
Unit of Measure	Percentage
Type of Measure: <ul style="list-style-type: none">• Assessment• Continuous• Discrete	Assessment only
Related dimension	Validity and Accuracy
Optionality	If a data item is mandatory, 100% completeness will be achieved, however validity and accuracy checks would need to be performed to determine if the data item has been completed correctly

Askham, N. et al, 2013, DAMA UK DQ Dimensions White Paper R3 7 available at <http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193&sid=6710711e8698efe5ed092128402a20b3>, downloaded on 20 December 2017

Completeness [2]

Example(s)	<p>Parents of new students at school are requested to complete a Data Collection Sheet which includes medical conditions and emergency contact details as well as confirming the name, address and date of birth of the student.</p> <p>Scenario:</p> <p>At the end of the first week of the Autumn term, data analysis was performed on the 'First Emergency Contact Telephone Number' data item in the Contact table.</p> <p>There are 300 students in the school and 294 out of a potential 300 records were populated, therefore $294/300 \times 100 = 98\%$ completeness has been achieved for this data item in the Contact table.</p>
Pseudo code	<p>Count 'First Emergency Contact Telephone Number' where not blank in the Contact table/ count all current students in the Contact table.</p>

Askham, N. et al, 2013, DAMA UK DQ Dimensions White Paper R3 7 available at <http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193&sid=6710711e8698efe5ed092128402a20b3>, downloaded on 20 December 2017

- See the work of Purdam and Elliot (2007) about the **reduction of analytical completeness** which relates to the fact that some disclosure control methods, typically the recoding of taxonomic schemes into coarser categorisations, mean that analyses that might have been conducted with unrecoded data cannot be

Validity [1]

Title	Validity
Definition	Data are valid if it conforms to the syntax (format, type, range) of its definition.
Reference	Database, metadata or documentation rules as to the allowable types (string, integer, floating point etc.), the format (length, number of digits etc.) and range (minimum, maximum or contained within a set of allowable values).
Measure	Comparison between the data and the metadata or documentation for the data item.
Scope	All data can typically be measured for Validity. Validity applies at the data item level and record level (for combinations of valid values).
Unit of Measure	Percentage of data items deemed Valid to Invalid.
Type of Measure: <ul style="list-style-type: none">• Assessment• Continuous• Discrete	Assessment, Continuous and Discrete
Related dimension	Accuracy, Completeness, Consistency and Uniqueness
Optionality	Mandatory
Applicability	

Askham, N. et al, 2013, DAMA UK DQ Dimensions White Paper R3 7 available at <http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193&sid=6710711e8698efe5ed092128402a20b3>, downloaded on 20 December 2017

Validity [2]

Example(s)	<p>Each class in a UK secondary school is allocated a class identifier; this consists of the 3 initials of the teacher plus a two digit year group number of the class. It is declared as AAA99 (3 Alpha characters and two numeric characters).</p> <p>Scenario 1: A new year 9 teacher, Sally Hearn (without a middle name) is appointed therefore there are only two initials. A decision must be made as to how to represent two initials or the rule will fail and the database will reject the class identifier of "SH09". It is decided that an additional character "Z" will be added to pad the letters to 3: "SZH09", however this could break the accuracy rule. A better solution would be to amend the database to accept 2 or 3 initials and 1 or 2 numbers.</p> <p>Scenario 2: The age at entry to a UK primary & junior school is captured on the form for school applications. This is entered into a database and checked that it is between 4 and 11. If it were captured on the form as 14 or N/A it would be rejected as invalid.</p>
Pseudo code	<p>Scenario 1: Evaluate that the Class Identifier is 2 or 3 letters a-z followed by 1 or 2 numbers 7 – 11.</p> <p>Scenario 2: Evaluate that the age is numeric and that it is greater than or equal to 4 and less than or equal to 11.</p>

Askham, N. et al, 2013,
DAMA UK DQ Dimensions
White Paper R3 7 available
at
<http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193&sid=6710711e8698efe5ed092128402a20b3>, downloaded
on 20 December 2017

- See again the work of Purdam and Elliot (2007) about the loss of analytical validity, which can be said to occur when a disclosure control method has changed a dataset to the point at which a user reaches a different conclusion from the same analysis

Accuracy [1]

Title	Accuracy
Definition	The degree to which data correctly describes the "real world" object or event being described.
Reference	Ideally the "real world" truth is established through primary research. However, as this is often not practical, it is common to use 3rd party reference data from sources which are deemed trustworthy and of the same chronology.
Measure	The degree to which the data mirrors the characteristics of the real world object or objects it represents.
Scope	Any "real world" object or objects that may be characterised or described by data, held as data item, record, data set or database.
Unit of Measure	The percentage of data entries that pass the data accuracy rules.
Type of Measure: <ul style="list-style-type: none">• Assessment• Continuous• Discrete	Assessment, e.g. primary research or reference against trusted data. Continuous Measurement, e.g. age of students derived from the relationship between the students' dates of birth and the current date. Discrete Measurement, e.g. date of birth recorded.
Related Dimension	Validity is a related dimension because, in order to be accurate, values must be valid, the right value and in the correct representation.
Optionality	Mandatory because - when inaccurate - data may not be fit for use.

Askham, N. et al, 2013, DAMA UK DQ Dimensions White Paper R3 7 available at <http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193&sid=6710711e8698efe5ed092128402a20b3>, downloaded on 20 December 2017

Accuracy [2]

Example(s)	<p>A European school is receiving applications for its annual September intake and requires students to be aged 5 before the 31st August of the intake year.</p> <p>In this scenario, the parent, a US Citizen, applying to a European school completes the Date of Birth (D.O.B) on the application form in the US date format, MM/DD/YYYY rather than the European DD/MM/YYYY format, causing the representation of days and months to be reversed.</p> <p>As a result, 09/08/YYYY really meant 08/09/YYYY causing the student to be accepted as the age of 5 on the 31st August in YYYY.</p> <p>The representation of the student's D.O.B.—whilst valid in its US context—means that in Europe the age was not derived correctly and the value recorded was consequently not accurate.</p>
Pseudo code	<p>$((\text{Count of accurate objects}) / (\text{Count of accurate objects} + \text{Counts of inaccurate objects})) \times 100$</p> <p>Example: $(\text{Count of children who applied aged 5 before August/YYYY}) / (\text{Count of children who applied aged 5 before August 31st YYYY} + \text{Count of children who applied aged 5 after August /YYYY and before December 31st/YYYY}) \times 100$</p>

Askham, N. et al, 2013, DAMA UK DQ Dimensions White Paper R3 7 available at <http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193&sid=6710711e8698efe5ed092128402a20b3>, downloaded on 20 December 2017

Consistency

Title	Consistency
Definition	The absence of difference, when comparing two or more representations of a thing against a definition.
Reference	Data item measured against itself or its counterpart in another data set or database.
Measure	Analysis of pattern and/or value frequency.
Scope	Assessment of things across multiple data sets and/or assessment of values or formats across data items, records, data sets and databases. Processes including: people based, automated, electronic or paper.
Unit of Measure	Percentage.
Type of Measure: <ul style="list-style-type: none">• Assessment• Continuous• Discrete	Assessment and Discrete.
Related Dimension(s)	Validity, Accuracy and Uniqueness
Optionality	It is possible to have consistency without validity or accuracy.
Example(s)	School admin: a student's date of birth has the same value and format in the school register as that stored within the Student database.
Pseudo code	Select count distinct on 'Date of Birth'

Askham, N. et al, 2013, DAMA UK DQ Dimensions White Paper R3 7 available at <http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193&sid=6710711e8698efe5ed092128402a20b3>, downloaded on 20 December 2017

Uniqueness [1]

Title	Uniqueness
Definition	No thing will be recorded more than once based upon how that thing is identified.
Reference	Data item measured against itself or its counterpart in another data set or database.
Measure	Analysis of the number of things as assessed in the 'real world' compared to the number of records of things in the data set. The real world number of things could be either determined from a different and perhaps more reliable data set or a relevant external comparator.
Scope	Measured against all records within a single data set
Unit of Measure	Percentage
Type of Measure: <ul style="list-style-type: none">• Assessment• Continuous• Discrete	Discrete
Related dimension	Consistency
Optionality	Dependent on circumstances

Askham, N. et al, 2013, DAMA UK DQ Dimensions White Paper R3 7 available at <http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193&sid=6710711e8698efe5ed092128402a20b3>, downloaded on 20 December 2017

Uniqueness [2]

Example(s)	A school has 120 current students and 380 former students (i.e. 500 in total) however; the Student database shows 520 different student records. This could include Fred Smith and Freddy Smith as separate records, despite there only being one student at the school named Fred Smith. This indicates a uniqueness of $500/520 \times 100 = 96.2\%$
Pseudo code	(Number of things in real world)/(Number of records describing different things)
External Validation	IAM Asset Information Quality Handbook Principles of Data Management, Keith Gordon

Askham, N. et al, 2013, DAMA UK DQ Dimensions White Paper R3 7 available at <http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193&sid=6710711e8698efe5ed092128402a20b3>, downloaded on 20 December 2017

Timeliness [1]

Title	Timeliness
Definition	The degree to which data represent reality from the required point in time.
Reference	The time the real world event being recorded occurred.
Measure	Time difference
Scope	Any data item, record, data set or database.
Unit of Measure	Time
Type of Measure: <ul style="list-style-type: none">• Assessment• Continuous• Discrete	Assessment and Continuous
Related dimension	Accuracy because it inevitably decays with time.

Askham, N. et al (2013, DAMA UK DQ Dimensions White Paper R3 7 available at <http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193&sid=6710711e8698efe5ed092128402a20b3>, downloaded on 20 December 2017

Timeliness [2]

Optionality	Optional dependent upon the needs of the business.
Example(s)	Tina Jones provides details of an updated emergency contact number on 1 st June 2013 which is then entered into the Student database by the admin team on 4 th June 2013. This indicates a delay of 3 days. This delay breaches the timeliness constraint as the service level agreement for changes is 2 days.
Pseudo code	Date emergency contact number entered in the Student database (4 th June 2013) minus the date provided (1 st June 2013) = a 3 Day delay.

Askham, N. et al (2013, DAMA UK DQ Dimensions White Paper R3 7 available at <http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193&sid=6710711e8698efe5ed092128402a20b3>, downloaded on 20 December 2017

Methods to deal with DQ

- On weeks 4-6 you will be looking at Methods to deal with the multiple components of data quality:
 - Data integration
 - Data description, summarization and visualization
 - Data cleaning
 - Data reduction
 - Data transformation
 - Data discretization and generalization

TAKE HOME MESSAGE

- Metadata is key to manage any type of data project
 - Data quality has to be addressed from the outset to create organised and efficient data infrastructures
 - Keeping track of how datasets are produced is key to maximise their use
- 