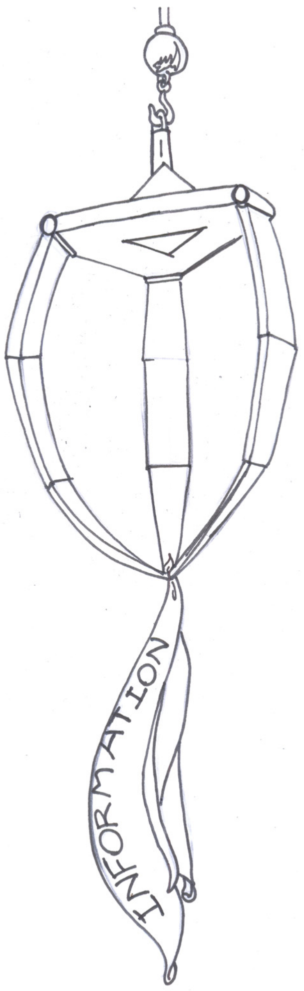
Chapter 6
Data Acquisition and Data Quality



No generalising beyond the data, no theory. No theory, no insight. And if no insight, why do research.

Henry Mintzberg

'Developing Theory About the Development of Theory,' in Ken G. Smith and Michael A. Hitt, *Great Minds in Management: the Theory of Process Development* (2005), 361.

One must never forget the dubious origin of the source data. The resultant data sometimes takes up a life of its own. Beware, no magic formula exists! (Reeve. 1997)

The current situation is one where research is suffering from **DRIPS**

Data-Rich-Information-Poor Syndrome[1]

Data is available in vast volumes – whether in book format or in digital real-time format. One must ensure that the data that is being gathered is reliable, has been gathered in the strictest manner and that it pertains to the topic under study and not to an alternative dataset.

One is never sure of the problems and issues pertaining to data gathering, however knowledge of what that data holds is necessary. Data has to be turned into information to avoid a DRIPS situation. Again the issue of context comes up. Information needs context so one needs to know which situational background that information was based on. If in default DRIPS occurs.

The next section gives an overview of which data forms exist and focuses on the topic of METADATA, which is the process of how one ensures that data has a context within which it was created and that it serves as a veritable ID card/Passport for that particular dataset.

Data Categories

As described in the previous chapters there are various types of data ranging from raw data to structured data, there is data that represents the function it was set out to do whilst others serve as surrogates. Then there is the pinnacle of data structuring; the metadata. Each is summarily described below:

i. Raw

The data that is gathered straight from the field and left in its pure form is called raw data. This category provides a stream of data that is either continuously or periodically gathered. It can be remotely gathered or from in-situ technologies. This type of dataset requires investment in cleaning and structuring.

ii. Numerics

This refers to data that has been given a context and can be readily analysed. The data can be structured in a way as to either represent a full population or a sample.

iii. Imagery

A category that is not normally associated with data can be found under the generic term imagery. Imagery is a multi-faceted mode that submits information to the user due to: its raw content, the position it is located in and the context it is located in. This data has been given a spatial context and can be easily understood in today's world of online maps, graphics and information systems.

A short pointer to a photo should do the trick. Take a photo as shown below (Figure 6.1) and look closely at the image.

[1] http://www.pharmamanufacturing.com/articles/2005/397.html

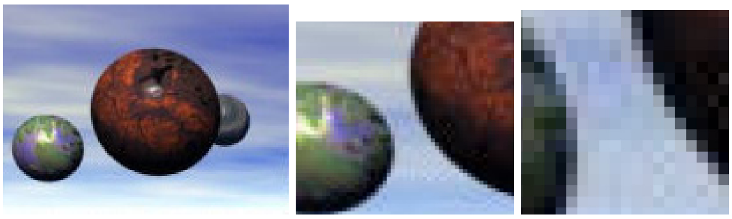
http://jobfunctions.bnet.com/abstract.aspx?docid=158118

Figure 6.1: Main Photo

The image should reveal itself as being composed of tiny dots, something also seen on a TV or monitor. Such structures are called pixels and each pixel depicts a colour, where the colour represents some kind of raw data. Each image can be composed of various collections of such pixels: for example an image of 800x600 is made up of 800 pixels wide by 600 pixels high, whilst a 1280x768 is composed of 1280 pixels wide by 768 pixels high. Quite a lot of pixels, one would say. The former image has 480,000 such pixels and the latter has 983,040 pixels. Best to leave a computing engine to calculate what data each should depict.

Now let us understand what each pixel is saying. Zooming in on a detail of the image shows that the spheres are composed of small pixels (Figure 6.2).

Figure 6.2: Pixels in an Image



Each colour (or greyscale) depicts various messages, amongst them:

- the boundary (shape of the object being depicted) – a round object that can represent a planet, a square boundary that represents a house, a polygon that can represent a field of corn;
- the containment of the pixel: the colour (represents either a direct colour or a false colour (example red would represent height, blue depth) – a blue sphere or a red one can indicate an Earth and a Mars;
- the adjacency (colour of the adjacent pixel which can describe what contained outside of the boundary) – the sky, background, the closest sphere.

Each pixel is given meaning from the above raw data through its interaction with the surroundings, again described by its context.

iv. Metadata

What is a Metadata?

One can immediately hear groans of frustration: isn't data complicated enough as it is without having to understand and learn about another level of data! Actually, metadata eases the understanding process impinged by data. It is simply data about data. A Metadata provides a description of what a dataset is composed of.

A metadata on an image might state the dimensions of an image: its width and height as well as the date and time it was taken. A metadata on a music file can describe the length of the composition, it's composer and a summary of the composition.

Whilst very detailed metadata documentation is available such as those described by International Standards (ISO 19115 and ISO 19119) under the INSPIRE[2] Directive[3] 2007/2/EC which outlined metadata specifications for spatial data, the scope here is to describe metadata in a simple form that would cover most data categories and not just specifically spatial data.

Basic elements that can be included in the metadata form include: information on the creator or guardian of the data, information on the data source and information on the contents of that dataset. Note that the metadata is not restricted simply to numeric datasets but also to imagery, documentation and any other archived and live forms.

A more detailed list would include the following items:

a. Who is the creator of the data? (the aim is to ensure that there is someone responsible for the maintenance and upgrading of that dataset)
   - i. Who owns it at the current date?
   - ii. Are contact details available?

b. Where did it come from? (aim is to source the original data)
   - i. Are details on the original dataset or document available?
   - ii. What scope was it created for?
   - iii. Is it updated periodically?
   - iv. When was it created in its present form?
   - v. When was the source data gathered

c. What does the data contain? (the aim is to help check whether one can repeat the capture as well as identify what is held within that dataset)
   - i. Title
   - ii. Format
   - iii. Attributes
   - iv. Medium such as dataset, database, spreadsheet, map, document, image
   - v. Spatial data such as scale, projection, coordinate system, bounds

d. Operational Issues (generic and overarching abstract data which would help persons querying the existence of access to a dataset.
   - i. Keywords
   - ii. Access issues
   - iii. Charges, if any
   - iv. Metadata on metadata

More detailed structures have been created such as the categories that are specified in the INSPIRE Directive Implementing Rules for Metadata, which inhabit a higher abstract level based on broad categories[4]. One can see that most are related to the spatial aspect, however one must not be derailed by the list in order not to create a metadata. If one identifies the data as non-spatial those related spatial elements can be switched off.

- Identification
- Classification of spatial data an services
- Keyword
- Geographic location
- Temporal reference
- Quality and validity
- Conformity
- Constraints related to access and use

[2] http://inspire.jrc.ec.europa.eu/

[3] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32008R1205:EN:NOT

[4] http://geostandards.geonovum.nl/index.php/6.6.3_INSPIRE_Metadata_elements

- Organisations responsible for the establishment, management, maintenance and distribution of spatial data sets and services
- Metadata on metadata

An online metadata creator based on INSPIRE can be found at http://www.inspire-geoportal.eu/index.cfm/pageid/342. It is straightforward to use and exports an output in xml format for further transposition to a discovery (search tool) service.

In terms of non-spatial data, the following metadata contains a list based on the INSPIRE output which can be used for non-spatial data. The following describes the broad categories of the metadata specified by the EU JRC INSPIRE Implementing Rules for Metadata[5].

- Identification
- Classification of spatial data an services
- Keyword
- Geographic location
- Temporal reference
- Quality and validity
- Conformity
- Constraints related to access and use
- Organisations responsible for the establishment, management, maintenance and distribution of spatial data sets and services
- Metadata on metadata

Always create a metadata for every datum created as it helps one to source back the relevant information and ascertain whether it is relevant for studies being carried out at considerable time post-creation. Together with a lineage, this tool helps one to ensure that the base data on which to run research is reliable, sourced and helps to ascertain whether one can use its attributes in real or surrogate forms.

Data Sourcing

Sourcing data for one's project is not a ready-made task. As technology is becoming more immersive, most organizations are creating their own datasets, allowing access either through online queries, or through a system of dedicated research units. Thus, theoretically, data sourcing is increasingly becoming transparent and easy to access. This said, one must look at the issues concerned with such sourcing.

Main questions to ask

Note that should any of these successive questions prove in the negative, then the options may be limited to a physical/manual check (where a dataset or part of it may be indicated) or to go back to reviewing the research scope. If the dataset is still required, then an actual collection by the researcher is necessary.

- Existence of a dataset
  - Does a dataset exist to one's knowledge?
  - Are there surrogates for the research/project purpose?

- Metadata
  - Is a full metadata available?
  - Does the metadata indicate that one can use such a dataset for the research/project?

[5] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32008R1205:EN:NOT

- Sources
    - Can one contact the sources?
    - Are the sources reliable?
    - Has one access to the sources?

- Availability
    - Is the dataset readily available?
    - Are there restrictions on access?
    - Is it protected by the Data Protection Act (of 2001)[6] and the Freedom of Information Act (of 2010)[7]?
    - Has one access to the actual dataset or to partial attributes?

- Costs and Access on time
    - Are there charges linked to the dataset?
    - Are these established on full recovery or are they based on a nominal charge?
    - Can the data be made available on time for the researcher to fit in the research/project?

- Fitness for researcher's needs
    - Once the data has been acquired, does it fit the real needs of the research?
    - Does the dataset reflect the metadata?
    - Can one go back to the originator in case the data does not fit the request?

- Acquisition
    - If the data is not made available has the researcher implemented a plan B for acquisition?
    - Can the user start the process to gather his/her own data?

Note that the acquisition of data from other sources may result in overkill or an under-representation of that data for the research needs. The data needs and interests of the creator are different from those of the researcher and as such that dataset may contain large volumes of extra data or even contain data that is not of the level required for the researcher's project.

As an example an inter-census (mid-point between the decadal census years) data request on housing stock is not available from statistics offices but can be acquired from such entities as a planning agency (which would have data on permits given but not necessarily on completions), from a utility-billing agency (which would have data on periodic metering but not on actual habitable dwellings) or from the postal office, which may have a list based on postcode but includes all types of other building types.

Each of the above can have missing or data that is not up to date, extra data such as non-dwelling units, amongst others. Therefore, the acquisition of pre-existing data limitations must be kept in mind. These include: the fact that the user has no control over the original format, over the content, over the attributes structure, over the actual data requested for as well as on the volume of data acquired.

One could even end up with very little data or a massively-populated dataset that includes extra data that is not required for the study but that the seller deems too cumbersome to dissect according to the query. Thus overpayment may occur at this stage. The best way forward is to ensure that the data costs reflect the study requirements and not to go overboard on acquisition. As stated earlier, data is very expensive and in certain cases costs more than the actual software and hardware used to analyse it.

Thus it is best to choose the datasets that fit the needs of the study and to ensure that data maximization occurs. The dataset acquired should also allow other users to gather once and use many times. This process ensures that more people can work on the same dataset and that they can add their own attributes, thus creating value-added datasets.

[6] http://www.dataprotection.gov.mt/

[7] http://docs.justice.gov.mt/lom/Legislation/English/Leg/VOL_16/chapt496.pdf

Data Capture

Data capture can occur through various modes which ensure that the process from data source to database is streamlined. This so-called data stream may employ manual or automatic methods of data capture. The automatic methods are governed by the rules integrated in the systems. Conversely, the manual ones may require: manual data input, data encoding, scanning, digitizing, and electronic data transfers. Knowledge of the different technologies is required, particularly knowledge of the software used and the availability of tools to help one become more efficient. These include OCR (Optical Character Recognition) technology that recognises text and converts it to digital text. More modern applications also allow dictation (speech to text) and calligraphy (handwriting) recognition.

Quality

Data quality is concerned with the creation of datasets that conform to specified requirements where the main emphasis is placed on prevention of error generation and not on post-creation analysis. To a certain extent, it is useless to analyse datasets for quality 'after' the whole process has been employed as against ensuring that the quality measures are put in place 'before' the capture even starts. An error generated at the start would be replicated throughout the capture and one may not have the chance at a second run. Other quality elements include the facts that nil defects must be generated during the capture, thus no room for error exists.

Error

Registering as hi-level quality issues, errors are nearly always generated during data capture, either through faulty processing, faulty technologies and sensors as well as data input mistakes.

Errors refer to the difference between the captured data and the real data that exists. This is sort of similar to the communication errors generated during interviewing when one person interprets a reply in a subjective way which may not be a true replica of what the interviewee meant it to be.

Errors can be categorised by type, by source, by the medium, by the technology and by the effects generated. Researchers should always be on the lookout for error generation. The following terms identify particular considerations that must be taken into account (Reeve, 1997):

- Accuracy – extent to which an estimated value approaches the true value
- Precision – level of recorded detail
- Scale and resolution – smallest size that can be displayed (for spatial datasets)
- Bias – systematic deviation from a norm or from the truth
- Completeness – extent to which data is supplied for all component parts and time periods
- Temporal consistency – repeated elements of the data handling process
- Logical consistency – suitability of commands, operations and analysis
- Semantic accuracy - quality with which objects are described
- Repeatability – extent to which independent users can produce the same data or output

Primary, Secondary and Tertiary Sourcing

There are 3 main classifications of data sourcing: primary, secondary and tertiary. Since the definition of data covers all data categories, this includes data types as defined in this book, inclusive of documentation.

Primary sources are those sources that point at data gathered first-hand. A thesis or a professional report for a company is considered as a primary source since it is composed of first-hand and unique information. This type of source also includes such first-hand information as: data gathered by other researchers (which is made available) as well as original documentation emanating from research or writing. This publication has covered such a source and termed it as 'raw data'. This data allows one to analyse at first-hand the original and unique data gathered and can run or rerun tests as well as compare to new data created by the researcher.

An example of a Primary Source: notes gathered during a field survey.

Secondary sources are those sources that are based on the findings of others such as those made available in academic journals. Researchers make references to these secondary sources in order to contrast and compare between different sources and then decide how those findings will affect their own study. Thus, secondary sources affect the outcome of how primary sources are captured.

An example of a Secondary Source: article written on notes gathered by a number of researchers during field surveys.

Tertiary sources are those sources that are not directly linked to an author or editor. These normally refer to actual data sources created by experts. A simple example of a tertiary source would be a book listing all the papers and books published on taxonomy. Papers will not be included, except for example abstracts. To a certain aspect, the academic search engines Athens, Science Direct and Ingenta all fit into the tertiary sources directory. The following is an example of a Tertiary Source: a list of titles and authors of all articles written on notes gathered by a number of researchers during field surveys and structured as an index.

The following amended General Classification (Table 6.1), sourced from the Department of Translation Studies, University of Tampere is a very useful tool to use during the differentiation of the type of source one may use.

Table 6.1: General Classifications of Selected Primary, Secondary and Tertiary Sources

Primary	Secondary	Tertiary
• Autobiographies • Correspondence • descriptions of travel • diaries • literary works • interviews • personal narratives • paintings and photographs • data gathered by researcher • data gathered through technological tools	• Biographies • prior books & papers on a topic • literary criticism & interpretation • history & historical criticism • political analyses • reviews of law and legislation • essays on morals and ethics analyses of social policy • study and teaching material • pre-prepared data	• Abstracts • bibliographies • chronologies • classifications • dictionaries & encyclopaedias • directories • guidebooks and manuals • population registers • statistics compendia

Source: Amended from: http://www.uta.fi/FAST/FIN/RESEARCH/sources.html

v. archival and real-time

Having covered quality, capture and sourcing, one other data issue that needs to be covered concerns the method of data capture. Data can be accessed depending on the sourcing. Primary data can be gathered in two main modes, that through archival and that through real-time capture.

Archival data is best referred to as original records that are gathered by the researcher and/or another researcher. This data is in its original state and has not been interpreted by others. The uniqueness issue is paramount and the archival records must be original records and the data gathered by the researcher must be his/her original work. All others fall under secondary or tertiary sourcing.

Capture Modes

One final item that has to be highlighted on data capture refers to the mode of capture. There are various ways through which one can capture data. These cover all the possible modes of capture and are not restricted to human-based capture.

Manual

This is the simplest and most commonly-employed mode. Data is gathered using time-honoured if not time-consuming effort. Researchers gather data manually through: field work, surveys, interviews and other methods as described in previous chapters. One needs to make use of analogue (hardcopies) material and then transpose findings into a storable system.

Semi-manual

This mode allows for the integration of tools together with the manual data sourcing. It includes the use of data capture technologies such as location-based maps that allow one to input data digitally into a coordinate system. This ensures that data is not inputted twice and spatial error is reduced drastically. Other tools, such as recorders (that aid the researcher in identifying keywords), fall into this category. Cameras also serve this purpose as an additional tool for data capture, particularly those enabled with face-recognition technology, voice-recognition and other innovative technologies that have become everyday tools for the researcher.

Automatic: in-situ/remote

These are the most complex and researcher-presence-free technologies. Automatic systems that gather information for the researcher are being employed more regularly in the real world and also in the virtual world. An example of the former would constitute someone who gathers data on air pollution (pm10) from an air monitoring station. In the virtual world case the researcher can employ electronic robots that gather data on users of particular sites such as the popular social networks examples of which are Facebook[8], Twitter[9] and Google Buzz[10].

In-situ automatic systems include such apparata as air monitoring stations, traffic cameras and council CCTVs. Remote apparata would include such items as drones (pilotless planes), satellites and airplanes.

In summary, data acquisition has to follow stringent rules to ensure that the data that is gathered can be transposed to information for eventual knowledge-building and action. Data acquisition depends on a plethora of capture issues such as: sourcing of new data, relevance to secondary sources, metadata structures, modes of capture and the reduction of error. Excluding any of these issues would endanger the outputs of one's study.

Questions (refer to Appendix for the answers)

1. Research is suffering from DRIPS. What is this condition and what should be done to avoid a DRIPS situation?

2. Briefly explain what metadata is.

3. List the three main data categories.

4. Why should a researcher always create a metadata for every datum?

[8] http://www.facebook.com/

[9] http://twitter.com/

[10] http://www.google.com/buzz

5. List the five main issues associated with the acquisition of pre-existing data.
6. What are research errors? When are they are mostly generated? How can they be categorized?
7. To avoid errors, researchers must consider certain factors. What are they?
8. List the three main classifications of data sourcing.
9. Primary data can be gathered in two main modes. Which are they?
10. What do you understand by "archival data"?
11. List the three main capture modes.