Reusable templates for the extraction of knowledge

Reusable templates for the extraction of knowledge

by

Paul J Palmer

A Doctoral Thesis

Submitted in partial fulfilment of the requirements for the award of
Doctor of Philosophy of Loughborough University

November 2020

Abstract

'Big Data' is typically noted to contain undesirable imperfections that are usually described using terminology such as 'messy', 'untidy' or 'ragged' requiring 'cleaning' as preparation for analysis. Once the data has been cleaned, a vast amount of literature exists exploring how best to proceed. The use of this pejorative terminology implies that it is *imperfect data* hindering analysis, rather than recognising that the encapsulated knowledge is presented in an inconvenient state for the chosen analytical tools, which in turn leads to a presumption about the unsuitability of desktop computers for this task. As there is no universally accepted definition of 'Big Data' this inconvenient starting state is described here as '*nascent* data' as it carries no baggage associated with popular usage. This leads to the primary research question: Can an empirical theory of the knowledge extraction process be developed that guides the creation of tools that gather, transform and analyse *nascent* data? A secondary pragmatic question follows naturally from the first: Will data stakeholders use these tools?

This thesis challenges the typical viewpoint and develops a theory of data with an underpinning mathematical representation that is used to describe the transformation of data through abstract states to facilitate manipulation and analysis. Starting from inconvenient '*nascent* data' which is seen here as the *true* start of the knowledge extraction process, data are transformed to two further abstract states: data *sensu lato* used to describe informally defined data; and data *sensu stricto*, where the data are all consistently defined, in a process which imbues data with properties that support manipulation and analysis.

The theory shows that when knowledge extraction is re-framed as a transformation of data state, the process may be implemented as reusable templates following the 'Literate Programming' paradigm. Working templates are demonstrated in both a synthetic example, and using real world data, to show how motivated end users can incorporate the outputs into their analytical workflow and to focus on the higher value interpretive aspects of analysis.

Acknowledgements

This work was funded under EPSRC grant: EP/R513088/1. The author also gratefully acknowledges contributions from the following organisations for access to staff and data in support of this research: The Leicestershire and Rutland Wildlife Trust; Rutland Water Nature Reserve (LRWT); Leicestershire and Rutland Environmental Records Centre (LRERC); NatureSpot; and the many volunteer county recorders responsible for collating and ensuring data accuracy.

Contents

List of Figures

List of Tables

List of Examples

Glossary

Abbreviations

BTO	British Trust for Ornithology
CLI	Command Line Interface
GUI	Graphical User Interface. Visually driven type of interface typically working in conjunction with a mouse or other pointing device.
GUI	Graphic User Interface
KPI	Key Performance Indicator. A measure used as a proxy for reporting effectiveness or performance.
LRWT	Leicestershire and Rutland Wildlife Trust
QDA	Qualitative Data analysis.
UCD	User Centric Design

Terminology

sensu lato (*s.s*)	Data that are defined in the loose sense.
sensu nascent (*s.n.*)	Data that are collated after the analytical tasks is defined.
sensu stricto (*s.l.*)	Data that are defined in the strict sense.
Antecedent data	Data that are collated before the analytical tasks are defined.
Long format data:	Narrow, or stacked data is presented with one column containing all the values and another column listing the context of the value. Not as human readable as Wide format data.

Wide format data: Wide, or unstacked data is presented with each different data variable in a separate column. Many tables presented as wide data for human readability.

Chapter 1

Introduction

Modern systems across many application domains can create huge amounts of data, often called 'Big Data', which can be kept almost indefinitely in low cost digital archives. This has led to the analysis of 'Big Data' becoming of practical and academic interest, as sources of such data proliferate and outgrow the analytic tools available (Sivarajah et al., 2017). 'Big Data' is typically noted to contain undesirable imperfections that are usually described using terminology such as 'messy', 'untidy' or 'ragged' requiring 'cleaning' as preparation for analysis. The use of this terminology implies that it is *imperfect data* hindering analysis, rather than recognising that the encapsulated knowledge is presented in an inconvenient state for the chosen analytical tools, which in turn leads to a presumption about the unsuitability of desktop computers for this task. As there is no universally accepted definition of 'Big Data' this inconvenient starting state is described here as '*nascent* data' as it carries no baggage associated with popular usage.

This thesis shows that keeping the physical manifestation of data separate from the abstractions of knowledge contained within, supports a perspective that challenges automatic assumptions that occur when the two viewpoints are mixed together. This leads to the primary research question: Can an empirical theory of the knowledge extraction process be developed that guides the creation of tools that gather, transform and analyse *nascent* data? A secondary pragmatic question follows naturally from the first: Will data stakeholders use these tools?

The value of answering this question is best seen from the stakeholder perspective, where the initial 'cleaning' phase is a manually intensive task lacking any supporting theoretical

structure. However, once the data has been transformed into a clean state, no matter how 'messy' and 'untidy' the source, a vast amount of literature exists exploring how best to proceed. The 'Tidy R' approach pioneered by Wickham (2014) is worthy of special note for the way in which it provides high level functions for simplifying the analysis by using a consistent rectangular metaphor for data. Thus, transforming stakeholder *nascent* data into a consistent rectangular format is key to eliminating the usual manually intensive 'cleaning' task.

This thesis challenges the established viewpoint and develops a theory of data with an underpinning mathematical representation that is used to describe the transformation of data through abstract states to facilitate manipulation and analysis. Starting from inconvenient '*nascent* data' which is seen here as the *true* start of the knowledge extraction process, data are transformed to two further abstract states: data *sensu lato* used to describe informally defined data; and data *sensu stricto*, where the data are all consistently defined, in a process which imbues data with properties that support manipulation and analysis.

The application of this theory demonstrates that by re-framing knowledge extraction as a transformation of data from an undesirable *nascent* state into the required *sensu stricto* state via an intermediate data *sensu lato* state, supports the creation of reusable templates which are an effective tool that motivated end users can incorporate into their analytical workflow, and that using templates in conjunction the 'Literate Programming' paradigm, allows users to focus on the higher value interpretive aspects of analysis.

The overall aim of this research is to find a way to extract knowledge from data by generalising the process to cover nascent data that is too large for manual inspection.

Specific objectives are:

- Constructing an underlying theory to model key aspects of the knowledge extraction process.
- Use the theory to implement a method that can prepare a wide variety of data for analysis.
- Evaluate the theory and implementation by working with data stakeholders and apply the templates to stakeholder directed analytical tasks

1.1 Overview Of Thesis Structure

This thesis is presented in eight chapters plus appendices summarised in Figure 1.1. Researching the current state of the art in this domain is a challenge in its own right as a search in Google Scholar for 'Big Data' returns more than five million entries. Thus, Chapter 2 'Literature review' requires the definition of a search methodology to explain how material was prioritised for inclusion in a short list for in-depth consideration. Details of the Literature review methodology are provided in Appendix A to support the summaries included in Chapter 2.

The methodology used to guide the search for an answer to the research question is described Chapter 3, and includes justification for the philosophical stance of critical realism adopted in this research. Viewing data as an empirical reflection of actual events in the real world leads to a realisation that user perspective and goals guide the analysis and interpretation of data, and suggests that a task orientated approach is a useful way to reduce the problem to its fundamental elements.

This work benefited from unrestricted access to real biodiversity data provided by the stakeholders, who are introduced in Chapter 4 'Data Stakeholders' along with justification for their selection in this research. These living datasets, were frequently updated, and the stakeholders found their analysis challenging because they consisted of multiple files, sources, and formats that have grown and changed over time. These real data were always found to exist in a disparate state, defined here as 'nascent data', not acknowledged in other academic work as the *typical* state of raw source data.

There are a number of programming languages that are candidates for building data analysis solutions, including R and Python. However, the obligatory use of the CLI, while efficient, requires technical skill and a certain mindset that inhibits adoption because of a user preference for GUI based software; so this work explores techniques to reduce these barriers. The visionary 'literate programming' concept of Knuth (1984), is still influential with the use of textile metaphors used to illustrate untangling the threads of: data; code; and narrative, to weave reproducible output documents. Textile metaphors such as 'knitting' and 'weaving' continue to influence the terminology around reproducible research and related software today. As an explanatory tool, a synthetic motivational example is introduced in Section 4.3 and used in call out boxes to illustrate technical details of problems and their

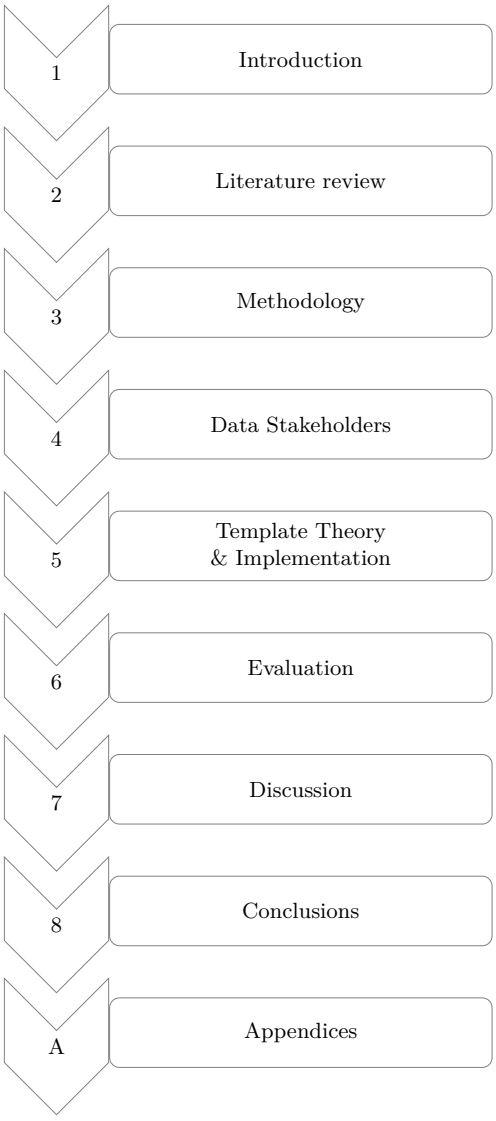


Figure 1.1: Overview of Thesis Structure

solutions. This example is synthetic in the sense that while all the specific challenges based on real world data, they are not necessarily all encountered at the same time.

The approach is described here, based on templates and the R analytical language, is compatible with the challenges of analysing Big Data on desktop computers. An underpinning theory for templates is developed in Chapter 5 which introduces two abstract concepts: *sensu lato* to describe informally defined data; and data *sensu stricto*, where the data are all consistently defined. An example of *sensu lato* is when data are accurate, but uses multiple formats for the capture of date and time information. This term is therefore distinct from 'untidy data' which is used to describe inconsistent data, as it is perfectly possible for data to be presented in a series of subsets, each of which are 'tidy', but collectively are data *sensu lato*. The theory of templates is independent of programming language unlike the implementation presented in Section 5.5, which is specific to the R language used in conjunction with R Studio and literate programming tools.

End users had a clear aspirational goal to learn 'better analytical techniques', and although they were highly motivated, they found achieving this goal difficult, with preparing data a particular problem. Chapter 6 evaluates the premise that using task orientated templates effectively reduces the coupling between user skill and analytical output by working with the data stakeholders previously introduced in Chapter 4. This work demonstrates a solution that uses reusable templates as an extension to the reproducible research concept and demonstrates how they may be used by end users to support their analytical tasks.

The broader potential of this template approach is discussed in Chapter 7 along with observations arising from the evaluation. For example, transformation of raw data is time-consuming to undertake by hand but is seen by users as an essential part of preparation prior to analysis. Using templates to prepare data provided considerable time savings for end users. By presenting the prepared data in a transportable format, such as csv, users were able to integrate it into their existing workflows providing instant benefits.

In addition, many analytical tasks are also repetitive and are an essential part of the interpretive process so may also be included within templates. Overall, this allows a much deeper analysis, so while the time savings in the analytical process are more limited, a greater knowledge capital is created. Finally, the conclusions are reviewed and summarised in Chapter 8.

Chapter 2

Literature review

This review explores the current state-of-the-art of extraction knowledge from large datasets, often called 'Big Data', with the ultimate goal of building an understanding of current capabilities. Because many application domains may be included in the Big Data umbrella, it is necessary to cast a wide net when seeking references that describe the current state of the art. This leads to a secondary challenge due to the magnitude of extant academic literature, such that it is not possible to review every publication manually. This has been addressed by the application of a strategy to filter the available corpus down to a reasonable size for critique. The method used was inspired by Sivarajah et al. (2017) in their review: 'Critical analysis of Big Data challenges and analytical methods' which had to address similar problems of filtering results to a manageable size. (Full details of the method are included in Appendix A). According to Sivarajah, the definition and classification of Big Data is yet to be fully established as it is a new and evolving discipline currently lacking in theoretical constructs. The context of this statement places the emphasis on *analysis* of Big Data, but seeking an understanding relating to the extraction knowledge directs attention to the nature of data, which is recognised here as the true starting point of this review.

A recurring theme in this thesis is the emphasis on how much Big Data analysis may be achieved with a desktop computer. While it might be argued that this is an arbitrary constraint, it helps to focus the mind on key issues within the overall domain where this research might have a more general impact, rather than confined to workers with privileged access to tools and data. Figure 2.1 gives an insight into topics uncovered within the initial

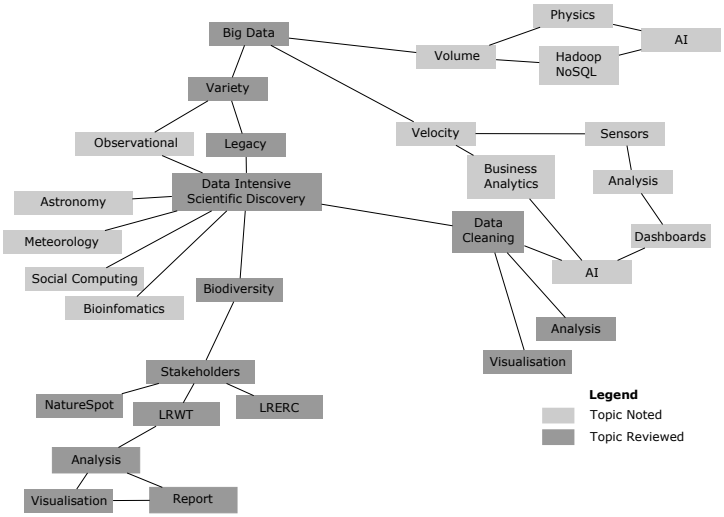


Figure 2.1: The literature review touched on many topics of interest, those shown with darker backgrounds were examined in more detail as they proved to be related to the research question.

scoping phase of this literature review. While only those shown with darker backgrounds were examined in more detail, it is clear that some topics have clearly defined boundaries, such as the high data volume physics, and the high velocity data of the business and sensor areas. Research contributions in these areas are likely to require access to both data and specialist hardware. Academic use of AI has become easier with the release of open-source libraries such as Google's TensorFlow (House of Lords Select Comittee, 2018), coupled with managed server access, would make this a very topical area to investigate deeper, but they were unlikely to contribute to the focus of this research.

Referring again to Figure 2.1, the 'Data Intensive Scientific Discovery' (DISD) paradigm is not predicated upon access to particular software or technology, but instead it seeks to incorporate shared data as a formal part of the scientific method (Tenopir et al., 2011). If DISD can be meaningfully undertaken on a desktop computer, then it is a much more accessible approach than one that requires specialist computational hardware. Thus, the focus on desktop computing is not a necessary condition for this research, however, as this work was undertaken on a laptop computer, the techniques should be replicable by any motivated researcher, a feature likely to improve impact.

The deeper examination of ecology shown on Figure 2.1, relates to the potential of DISD to the biodiversity community noted by multiple authors including Madin et al. (2007); Hochachka et al. (2012); Kelling et al. (2009). This observation led to the identification of stakeholders prepared to share data supporting this research and confirmed a preference for local analysis running on desktop computers.

2.1 Structure Of This Review

This review is structured as follows:

- Section 2.2 Defining the domain of interest: A description of the areas covered by this review; Areas excluded; Search methodology; Related reviews.
- Section 2.4 The size of Big Data: Key research in this domain; Interpreting the literature.
- Section 2.9 Questions: Where to next.

2.2 Defining The Domain Of Interest

This review explores the current state-of-the-art with regard to extraction of knowledge from large datasets, but before we can embark on this task, it is first necessary to define exactly what is meant by data, before we qualify it with an adjective 'big'. The Oxford English Dictionary (OED Online., 2020) defines data as:

> Related items of (chiefly numerical) information considered collectively, typically obtained by scientific work and used for reference, analysis, or calculation.

While this definition is both correct and authoritative from a language perspective, it does not help with the technical complexities encountered when analysing data with computers. Fortunately, a useful technical definition was produced by Fox and Levitin (1994) while exploring the concept of data quality in computerised databases. Based upon the work of Tsichritzis, Dionysios C and Lochovsky (1982), Fox noted that defining data as a collection of datum triples: entity, attribute and value provides for a versatile method of linking real world observations to representations of data that includes the data model along with the data representation as an integral part of the concept. Figure 2.2 is redrawn from Figures 1 and 2 of the 1994 reference to show the relationship between components that comprise this view of data. The terminology and concepts are used in the following section to help understand the many definitions of 'Big Data' found in the literature.

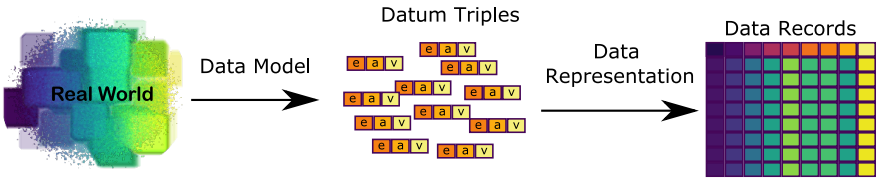


Figure 2.2: Data defined as a collection of datum triples: entity, attribute and value, linking real world observations to data records. This is loosely indicated by the choice of colour swatches; the Real World observations are arranged as rows of entity values, with attributes as the column headers to form the data records. Diagram redrawn from Fox and Levitin (1994).

2.3 Defining Big Data

We now move to the task of defining Big Data and begin with noting that there is no universally agreed definition. One author alone, Fosso Wamba et al. (2015), cite ten plausible definitions, further indicating the difficulty of pinning down what all agree is an important topic. Also, many authors implicitly choose to examine Big Data from a single perspective; for example, Elgendy and Elragal (2016) state without explanation, that dataset size is the most important characteristic of Big Data. In a similar context, frequently used without supporting references are the: 'n V's of Big Data': Volume, Variety, Velocity, Value, Veracity etc. While this terminology fits in with the popular narrative regarding big data, the lack of formal definitions limits its usefulness.

A systematic approach to defining the scope of Big Data has been attempted by De Mauro et al. (2016a) who found that papers relating to Big Data could be classified into four themes: Information, Technology, Methods, Impacts, but the chosen terminology does not sit well with Figure 2.2. However, renaming the groupings allows an informative mapping of the groups, as shown in Figure 2.3, which clearly shows that most definitions relate to aspects of data representation and records rather than the overall nature of data. The curated list of definitions is presented here before providing an interpretation:

Group A. Definitions based upon attributes of data representation.

1. Big Data requires a revolutionary step forward from traditional data analysis, characterized by its three main components: variety, velocity and volume (Sagiroglu and Sinanc, 2013).

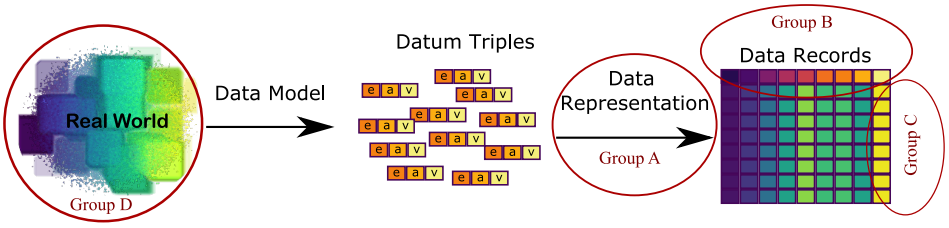


Figure 2.3: The groups of Big Data definitions from the main text are mapped onto the overall definition of data to illustrate the current academic focus.

2. The four characteristics defining big data are: volume, velocity, variety and value (Dijcks, 2013).

3. High volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making (Beyer and Laney, 2012).

4. Complex, unstructured, or large amounts of data (Intel IT Center, 2014).

5. Big data is a combination of Volume, Variety, Velocity and Veracity that creates an opportunity for organizations to gain competitive advantage in today's digitized marketplace (De Mauro et al., 2016a).

6. Can be defined using three data characteristics: Cardinality, Continuity and Complexity (Suthaharan, 2014).

7. Big Data: data captured from sensors, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, etc. (IBM, 2018).

8. Data from everything including click stream data from the Web to genomic and proteomic data from biological research and medicine (Davenport, 2012).

9. Big Data has three main characteristics of Big Data: the data itself, the analytics of the data, and the presentation of the results of the analytics. Then there are the products and services that can be wrapped around one or all of these Big Data elements (Gantz and Reinsel, 2012).

Group B. Definitions based upon the physical format of data records.

10. The storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning (Ward and Barker, 2013).

11. The process of applying serious computing power, the latest in machine learning and artificial intelligence, to seriously massive and often highly complex sets of information (Microsoft, 2013).

12. Extensive datasets, primarily in the characteristics of volume, velocity and/or variety, that require a scalable architecture for efficient storage, manipulation, and analysis (Big Data Public Working Group, 2018).

C. Definitions based upon numbers of data records.

13. A dataset that is too big to fit on a screen (Shneiderman, 2008).

14. Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse (Manyika et al., 2011).

15. Data that cannot be handled and processed in a straightforward manner (Fisher et al., 2012).

16. The data sets and analytical techniques in applications that are so large and complex that they require advanced and unique data storage, management, analysis, and visualization technologies Chen et al. (2013).

17. Data that exceeds the processing capacity of conventional database systems (Dumbill, 2013).

18. Data that you cannot load into your computer's working memory (Havens et al., 2012).

19. For statisticians, it is simply when there are numerical problems in linear algebra on large dense matrices (Bivand and Krivoruchko, 2018).

20. For GIS users, it is when there is a problem with data storage and data querying, usually less than one hundred thousand points (Bivand and Krivoruchko, 2018).

21. Extremely large sets of data related to consumer behaviour, social network posts, geotagging, sensor outputs (IBM, 2018).

D. Definitions linking to 'Real World' observations.

22. A cultural, technological, and scholarly phenomenon that rests on the interplay of Technology, Analysis and Mythology (Boyd and Crawford, 2014).

23. Phenomenon that brings three key shifts in the way we analyse information that transform how we understand and organize society: 1. More data, 2. Messier (incomplete) data, 3. Correlation overtakes causality (Mayer-Schonberger and Cukier, 2013).

The incomplete coverage of the definitions in Figure 2.2 supports a comment made by Sivarajah et al. (2017) that there is:
' ...a distinct lack of theoretical constructs and academic rigour ...in the field of big data research.'
This is not to suggest that Big Data is a term without meaning or it is used to describe a research area without merit, but more that real understanding is still being developed. A notable observation is the disconnect between big data definitions and the real world, as this suggests that there may be opportunities to improve understanding by considering the complete data journey. From this viewpoint, there is no convincing reason why Big Data (albeit that it is a useful umbrella term) is any different from other data and that the definition by (Fisher et al., 2012) is suitable when used with an explanatory caveat to qualify a problem:

> Big Data is that which cannot be handled and processed in a straightforward manner.

Collectively these definitions have arisen out of the need to describe the focus of academic challenges, but it is clear that relatively little progress has been made in understanding how to maintain the connection between data and its real world interpretation. This is phrased succinctly by Kitchin *et al* Kitchin et al. (2015):

> ...A critical understanding of data recognizes that data do not exist independently of the ideas, instruments, practices, contexts, knowledges and systems used to generate, process and analyse them, regardless of them being often presented in this manner....

This statement implies that one should always look at data in the context of our world view, rather than in isolation, that is to say, there is always a viewpoint that is implicitly applied to our interpretation of data. This is not saying that measurable attributes do not matter, but rather, they are just one aspect in a complex interplay between data and interpretation. Put another way, each of the above definitions for Big Data are valid in a particular context, but the relationship between 'raw data' used in research and the real world are often left incompletely described (Bowker et al., 2013) limiting the possibility of reproducibility.

2.4 The Size Of Big Data

The preceding section proposed adopting a definition of convenience which leads to an obvious and frequently asked question based on exceeding thresholds which is: 'how big is Big Data?' To explore the impact of dataset size on calculation time with a desktop computer, Jacobs (2009) presents an analysis of the Big Data challenges due to scale using US census data justifying the choice because most data has temporal and geospatial dimensions of the type captured in this set. Using this as a test case a number of pertinent observations are made:

- As relational databases become larger, the time to extract results becomes longer than for non-normalised tabular databases, however, the storage size occupied by them is much larger than their relational equivalent.
- Computational time approximates to an $Nlog(N)$ growth with the number of data rows N, so the calculation time will take too long to be useful as the number of rows grows too large.

The simplified implications are that for big data, large rectangular (tabular) databases, as shown on the right-hand side of Figure 2.2, are faster to access and that number of rows are a proxy for computational time. While these observations do not set quantitative limits for big data, due to combinations of factors, including specification of the computer used, they do help with interpretation. Considering a recent topical example is helpful in that it sets a contextual viewpoint for big data and its relationship to the real world.

The so called 'Cambridge Analytica scandal' relates to the scraping of 87 million Facebook user profiles (Schneble et al., 2018; Editorial, 2018; Chang, 2018) and the manner in which that data was subsequently used. The tone of these references leaves the reader in no doubt that, because of the data size involved, this episode can be framed as a 'Big Data' issue. No sources cite the physical size, but if we compare against the UK Companies House basic dataset (BEIS, 2018) which is publicly available, we can make a plausible estimate. This dataset comprises 4,000,000 rows of 37 fields and is 2.04GB as an uncompressed text file. The data size per row is 0.5KB. Returning to the Cambridge Analytica data, if a similar conversion factor is used, then the complete dataset will be 43.5GB. If true, the dataset would easily fit on an ordinary USB drive. The current limit for Microsoft Excel spreadsheet is 1,048,576 rows by 16,384 columns Microsoft (2020). Thus, for Excel users, the Cambridge Analytica data is inaccessible Big Data. However, using the R analytical language, the author has trivially read a much larger geospatial dataset with a similar conversion factor comprising 30 million rows, suggesting that at just three times larger, the Cambridge Analytica data is plausibly within analytical range of a desktop computer.

This example highlights the choice of software tools as a factor in the perception of what might be considered as Big Data and also why knowledge of the content and real world context are required to assign commercial value. Returning to Figure 2.2 we can see that the real world of Facebook users are mapped through the data model of user profiles into a huge number of digital datum elements ready for transformation into data records, and analysis. It is suggested by the author that in this example, commercial value lay not just in data size, but in the clear and well-defined path between the real world and data records that enabled the extraction of valuable knowledge. It follows more generally, that seeking techniques which can help preserve this connection are an asset in the analytical process.

The challenges and commercial opportunities of Big Data are now widely recognised, and Manyika et al. (2011) estimated the associated value to the American and European economies were in excess of $1 trillion. Moving on a few years Cavanillas et al. (2016), quotes an estimate of €16 trillion by 2020. These values support the references to exponential growth in value of Big Data, and suggest that there is implied commercial worth in the tools to extract actionable knowledge from data. The value is said to be associated with innovation, competition, and productivity, but a cautionary tone is used here to reflect the lack of detail in these claims. There is a strong connection between Big Data and AI in the

commentary relating to the realisation of this value, a point emphasised by in the report from the Royal Society (2017). Access to 'Open Data' and the adoption of standards to support are Machine Learning (ML) are seen as important for the UK to participate in this area, but as the demonstrated by the discussion of the 'Cambridge Analytica scandal', there are areas which, while beyond the range of popular software, might be usefully analysed with desktop computers. See Figure 2.4.

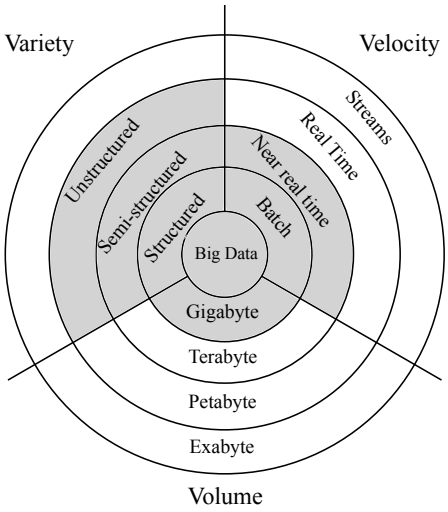


Figure 2.4: Three V's of Big Data. (*After Sagiroglu* Sagiroglu and Sinanc (2013).) Shaded areas are those considered to be within range of desktop computers in this work.

2.5 Big Data Analysis Tools

As suggested by the groupings of definitions from Figure 2.2, there is a focus on analysis of data based on attributes, format and size, which is reflected in the choice of tools chosen by respondents who identify themselves as data scientists typically working with Big Data. For example, Akiwatkar (2017) reports the following applications are used, and includes an indication of popularity:

1. R 48%.
2. Python 45%.
3. SQL 35%.
4. Excel 32%.

This survey is in line with the findings of Ali et al. (2016) in their review of 'Big data visualization: Tools and challenges', which also notes the popular use of cloud tools such as Tableau, Microsoft Power BI, and Plot.ly. A point made by Ali *et al* is that all of these tools can all be used in conjunction with R, Python and SQL for the pre-processing of Big Data prior to uploading and analysis, *if users have the necessary skills*, which is interpreted here as an oblique reference recognising the importance of applying the data definition of Figure 2.2, typically described as 'data cleaning.' The framing of programmatic tools as both difficult to use, and only for use in pre-processing leads the reader in easy steps to a viewpoint where it is problems with data that limit the application of higher level tools, rather than limitations in the tools themselves. Using the flexibility of programmatic tools to correct these problems is presented as a precursor step, which we now examine in more detail.

The Python and R languages both have well-supported libraries for data-analysis and visualisation in which many authors such as: Zhao (2015); Layton (2015); Slater et al. (2016); Ognyanova (2016); Hill and Scott (2017); Kandel et al. (2011); Wickham (2014); Kulkarni and Takawale (2015) and Cohen et al. (2009), describe normative approaches to data analysis and visualisation. Typically, SQL style commands are used to interrogate database servers to extract subsets of data, and in the Big Data context it may also be used to filter data into smaller, more manageable datasets (Russom, 2011). The current version of Microsoft Excel

is limited to 1,048,576 rows (Microsoft, 2020), so while a very useful application, cannot be regarded as a primary Big Data desktop analysis tool. However, it is a *de facto* calculation tool of choice for business, so its presence in this list is not unexpected. The tabular format of spreadsheet data was noted in Section 2.4 to be appropriate for Big Data, so it is not surprising that the popular 'Tidy R' approach of Wickham and Grolemund (2016) uses the same conceptual tabular format in conjunction with 'verbs' to manipulate data too large to be viewed as a spreadsheet.

Normative papers describing data mining techniques contain an implied connection to Big Data; Zhao (2015) provides examples of many excellent data mining techniques using the R language and briefly touches on issues around working with limited memory. Indeed, the data.table class in R has capabilities that make it much faster than other similar classes when used with very large datasets (R Core Team, 2017), but as it is a more difficult construct to use is not covered in introductory texts and so is less well known. While exploring the potential of Big Data for the creation of official statistics Daas et al. (2015) used a Windows 7 workstation running R and their case studies noted the need to clean the data prior to use and the need to address the statistical challenges of: coverage, representativeness, quality, accuracy and precision. It is clear that the intent of official statistics, and all the data mining examples, are to make accurate, quantitative observations about the real world, and that achieving this goal is limited by problems within data. While, this observation is true, attention is again drawn back to Figure 2.2 and that connecting data to the real world is more than data representation and records. It is suggested that the data model used to map the real world into datum triples has an overlooked role in the overall analytical process. This is perpetuated by a starting assumption, made in most of the work covered above without discussion or qualification, regarding the need to 'clean' source data prior to analysis, rather than looking for methods to work with data in the form in which it is presented.

2.6 Analytical Techniques

No reviews have been found that explicitly focus on the challenges associated with desktop analysis of Big Data. Informative discussion about analytical techniques are often agnostic about the technological implementation, for example, analytical techniques requiring the

representation and manipulation of information as matrices are relevant over much longer timescales than the underlying enabling computer technologies that use them. The paper by Andor et al. (1985) on the use of Galois Lattices to represent knowledge, remains a useful primer to the topic which has application in conjunction with Big Data.[1] Another important analytical tool, often used in the Big Data context, is clustering which may be defined as: the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The paper by Jain et al. (1999) contains some cautionary notes about applying clustering algorithms to Big Data[2] and gives examples of how misleading results may occur.

These older papers highlight the weakness of using a purely keyword related search strategy when trying to uncover papers describing abstract analytical techniques for Big data as they were written before the term was in common use. Unless the paper is applied explicitly to a Big Data problem, it is unlikely to be ranked highly in a search. As a specific example: Tumminello et al. (2011) 'Statistically Validated Networks in Bipartite Complex Systems' do not use any keywords that would ordinarily be associated with Big Data. However, the authors apply a statistical validation methodology to three networks: organisms; financial stocks; and films from the Internet Movie Database. Each of these cases meets at least one of the definitions of Big Data in Section 2.3. The author concludes that network representation of complex systems is highly relevant to this work, with a rich ecosystem of academic research. However, because of its underpinning nature, with applications across many domains, relevant publications are unlikely to be found using simple 'Big Data' related keyword searches. This search bias has been mitigated by carefully reviewing cited references within documents that have been reviewed in detail.

More generally, the statistics community recognise that working with Big Data introduces new type of statistical challenge in part because Big Data are not usually collected as part of a designed methodology with parameters selected for the testing of a hypothesis. 'Collect data first, ask questions later' is a more likely approach (Quarteroni, 2018), so knowledge discovery is more exploratory in nature, alluding to the Data Intensive Science paradigm. Ceri (2018) calls for a combined approach from statisticians and computer science to make progress on these challenges and comments:

[1] Using Google Scholar, this paper has been cited by 14,474 other papers. The keywords: 'Big Data Galois Lattice' returns 5,900 hits, many of which link to recent publications.

[2] The terminology of the day was Data Mining.

> ...in the past the statistical community has committed to the almost exclusive use of stochastic models....

This statement refers to the intentional isolation of statistical models from the real world to enable their mathematical justification. Unfortunately, in the case of big data, there may be many patterns uncovered that are of potential interest, but finding and applying the correct statistical models to validate those which are relevant, rather than those which are random, may be very difficult (Wong, 2018). An example of this is provided by Hochachka et al. (2012) in their paper 'Data-intensive science applied to broad-scale citizen science'. Data quality and interpretation are key issues addressed in the work which demonstrates how carefully applied statistical models can provide useful biodiversity insights. However, they also note that:

> ...once the data are collected and passed through quality control processes, we have found that existing methods for analysis may not always be suitable for use with such broad-scale observational data. ...

This remark provides support for looking beyond the data to include improved analytical techniques in the search for progress in this area. This remark highlights the need for developing improved analytical techniques for interpreting data collated outside controlled experimental conditions. Causal inference models are an area of particular interest that may address this challenge. Pioneered by Pearl (1995), causal diagrams allow the incorporation of contextual knowledge into statistical analysis of data. More recent work (Pearl, 2020) demonstrates the value of the approach in multiple domains including Big Data, and how it may be applied to provide robustness against missing data (Mohan and Pearl, 2018). Returning again to the terminology of Figure 2.2, causal diagrams may be thought of as representing the contextual connection between the real world and the data records. That real world context is needed for data interpretation has already been made, and here we see that it is also possible to embed it within a statistically justified analysis, outside experimentally controlled conditions.

Coupled with the application of appropriate statistical models, is the need to ensure that research is underpinned with reproducible results (Candes, 2017), which may be achieved by publishing the data and transformations applied to it alongside publications. Indeed, requiring authors to publish data for independent verification is now a requirement of some

high ranking journals, and is a trend that is only expected to increase (Callaghan et al., 2012; Tenopir et al., 2011; Parsons et al., 2010; Lawrence et al., 2011). These questions around the reproducibility of research have raised interest in techniques that combine report narrative, analytic code, and data into a single transparent process (Gentleman and Temple Lang, 2007). Stodden et al. (2014) focusses on the importance of good statistical practice when analysing data and cautions that all calculations should be presented with a route back to original data, so results may be verified. In the UK, the main funding bodies now require data created during publicly funded research to be made available for future analysis e.g. (EPSRC, 2014; IEEE, 2018; Callaghan et al., 2012). The UK Data service[3] is one of a growing number of research council recognised repositories that may be used for the sharing of data. There is also a trend towards government data to be shared on an open licence, but current research shows that the goals of availability and quality of such data are not always met (Vetrò et al., 2016). Sharing the code used to analyse data is less common but is already done within some scientific communities.[4]

Gentleman and Temple Lang (2007) went a step further and proposed the concept of a compendium type container for one or more dynamic documents and the different elements needed when processing them, such as code and data. The compendium then becomes a means of sharing all aspects of the research as a reproducible unit. Searching has not uncovered any evidence to suggest that this interesting idea was ever directly developed further by this author. However, Gentleman is credited for his work in the creation of the R analysis language (R Core Team, 2020) which is a key element in the practical implementation of dynamic documents described in Chapter 5 so it seems likely that he influenced the development of R language tools that support reproducibility which are described below.

The starting point for this exploration is the visionary literate programming work of Knuth (1984)[5], which used textile metaphors to illustrate the untangling of threads of data, code and narrative, to weave reproducible documents. Textile metaphors such as 'knitting' and 'weaving' continue to influence the terminology around reproducible research and related software today. A well-developed and supported suite of software for the production of reproducible documents is provided by R Studio (RStudio Team, 2016), Markdown (Cone, 2018), Pandoc (Macfarlane, 2017), LaTeX (Latex Team, 2020) and the optional Bookdown

[3] `https://discover.ukdataservice.ac.uk/`
[4] The Astrophysics Source Code Library `http://www.ascl.net/` is a notable exception.
[5] Robert Gentleman cited this reference with the incorrect date of 1992.

package (Xie, 2017). Markdown is an easily learned markup syntax used in conjunction with R Studio IDE that enables document structure, narrative, and R code to be inserted within a text file, conventionally with a `Rmd` extension to denote the internal format. This supports a natural workflow for interacting R, by executing short stanzas of R code, building the analysis as a series of small steps with intervening narrative. At anytime the complete document can be woven into a completed output document, usually PDF, Word, or HTML. In conjunction with Bookdown, publication ready book style PDF, HTML and epub formats are also possible. The Rrticles package (Allaire et al., 2019) supports the production of output in the required format for many scientific journals. Given the versatility of this open-source software ecosystem it is surprising that it is not more widely known. The author can vouch from personal experience that markdown presents an easier interface for the writing of technical documents than LaTeX, which was used to produce this thesis. Any edits to the output are made and the sub-files recompiled to build a new output document. This is a versatile approach which offers a high degree of integration between program and narrative. However, as with all dynamic documents, it requires good programming skills to make it work well, which is probably the most obvious barrier to its more widespread adoption.

Other mature systems exist for the creation of dynamic documents following literate programming principles: the Jupyter Notebook is an open-source web application for sharing documents with live code, data and narrative. Although the application interface is browser based, it can also be used on a desktop computer with a web server running on localhost (Yu et al., 2017b). Initiating local access to a Jupyter notebook and server requires access to a command line interface. However, Jupyter notebooks are already an established way of sharing code and data within the astrophysics community. An example of this is the first confirmed detection of gravitational waves from a black hole merger Abbott et al. (2016). Jupyter notebooks running on Microsoft Azure cloud computing service containing the code and data are freely shared (LIGO Scientific Collaboration, 2016).

The overall differences between R Studio and Jupyter Notebook dynamic document ecosystems are subtle and the research approaches in the following chapters could have been developed on either system, leaving open an element of personal choice. Perhaps the principle feature to consider is for users requiring compatibility with programming languages

other than R, the Jupyter Notebook is an attractive solution for producing dynamic documents. Users requiring a sophisticated IDE for R will find R Studio meets their needs. This research required support for dynamic documents and would have succeeded with either solution.

A notable gap in the area of reproducible documents is that no research has been found that reviews the security and integrity of code and data from mischievous actors in the context of open access academic publication and reproducibility even though the necessary tools are in common use. For example, many commercial and open source executable code and scripts are digitally signed to verify provenance (Microsoft, 2018; GnuPG, 2013), and in principle, a similar system could be extended to ensure that code and data shared for research has not been altered (The CA / Browser Forum, 2011). Methods for implementation of digital keys for cloud sourced data have been proposed by Dongare and Kadroli (2017) and include the additional sophistication of key revocation. If data are to be shared and reused the importance of such signing systems cannot be understated.

2.7 Data Visualisation

The visualisation of data is an important step in the communication of analytical results, and a single paper, 'The eyes have it: a task by data type taxonomy for information visualizations' by Shneiderman (1996) is probably the most cited paper relating to visualisation of Big Data, due in part because of the catch phrase:

> 'Overview first, zoom and filter, then details-on-demand',

repeated 10 times in succession to emphasise the author's point [6]. The paper remains relevant to the present day due to its insights into what has become the Big Data and the broad applicability of the data classifications it proposes. The more recent paper by Wang et al. (2015) presents a more up to date review of techniques that are now available when the constraints of Big Data are applied. Their list of data visualisation myths, shown as bold text, are a useful extension of Shneiderman's mantra:

> **All data must be visualised:** It is important not to overly rely on visualisa-

[6] Google Scholar lists 5009 citations of this paper.

Type	Description	1-D	2-D	3-D	Temporal	Manifold	Tree	Network
Scatter plot	Used to suggest correlations in data	No	Yes	Yes	Yes	Yes	No	No
Box plot	Check distribution of data against statistically normal values	Yes	Yes	No	Yes	No	No	No
Bubble chart	Similar to scatter chart where size / colour/ shape point used to represent additional dimensions	No	Yes	Yes	Yes	Yes	No	No
Line chart	Used to plot ordered series of values	No	Yes	Yes	Yes	No	No	No
Area chart	Similar to line chart, but area under the line is used to represent volume	No	Yes	Yes	Yes	No	No	No
Streamgraph	Type of stacked area graph	No	Yes	Yes	Yes	No	No	No
Pie chart	Illustrates relative proportions as segments of a circle	Yes	No	No	Yes	No	No	No
Histogram	Representation of the distribution or numerical data	Yes	Yes	No	Yes	No	No	No
Density plot	Representation of the distribution or numerical data	Yes	Yes	No	Yes	No	No	No
Heat map	Displays individual values contained in a matrix as colours	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 2.1: Quantitative visualisation charts and their applicability for data classified with Shneiderman's data taxonomy (Shneiderman, 1996).

tion; some data does not need visualisation methods to uncover its messages.

Only good data should be visualised: A simple and quick visualisation can highlight something wrong with data just as it helps uncover interesting trends.

Visualisation always leads to the right decision: Visualisation cannot replace critical thinking.

Visualisation will lead to certainty: Just because data is visualised doesn't mean it shows an accurate picture of what is important. Visualisation can be manipulated with different effects.

No comprehensive review of data visualisations has been found, so a summary has been created for this report by noting the general usage in papers and software manuals. Tables 2.1, 2.2, and 2.3, list a range of data visualisations with their potential use classified with Shneiderman's data taxonomy (Shneiderman, 1996). The Scatter and Box plot are often recommended as initial checks when exploring data as they provide information on the nature

Type	Description	1-D	2-D	3-D	Temporal	Manifold	Tree	Network
Bar chart	Used especially for categorical data	Yes	No	No	Yes	No	No	No
Time line	Used to represent categorical values along a time axis	No	Yes	No	Yes	Yes	No	Yes
Gantt chart	A stylised representation of project task and their temporal relationships	No	Yes	No	Yes	Yes	No	Yes
Flow chart	Depicts a process or other system	Yes	Yes	No	Yes	Yes	Yes	Yes
Parallel coordinates	Used to visualise high dimensional geometry	Yes	No	No	No	Yes	No	No
Tree map	Display hierarchical data as nested elements	Yes	No	No	No	Yes	Yes	Yes
Dendrogram	Displays clusters by hierarchical discrimination	Yes	No	No	No	Yes	Yes	Yes
Semantic network	Network diagram capturing the relationship between concepts	No	No	No	Nol	Yes	Yes	Yes

Table 2.2: Categorical visualisation charts and their applicability for data classified with Shneiderman's data taxonomy (Shneiderman, 1996).

Type	Description	1-D	2-D	3-D	Temporal	Manifold	Tree	Network
Geospatial map	Map projection using a geospatial coordinate system	No	Yes	Yes	No	No	No	No
Voroni Diagram	Divides a plane into regions in relation to predefined points that act as seeds	No	Yes	Yes	No	Yes	No	No
Choropleth map	Displays changes in a variable on a map by region	No	Yes	Yes	No	Yes	No	No
Cartogram	Map in which a thematic mapping variable such as population is substituted for land area or distance	No	Yes	Yes	No	Yes	No	No

Table 2.3: Geospatial visualisation charts and their applicability for data classified with Shneiderman's data taxonomy (Shneiderman, 1996).

data that may help guide the analytical process (Zhao, 2015). All types of chart may be used in conjunction with faceting, the generation of a series of related charts along a dimension of interest, to show progressive changes e.g. A series of maps showing changes in population density over time. It is also possible to overlay charts of different types. For example, weather forecasts routinely overlay temperature, wind and rainfall over a geospatial map to highlight regional variations. These maps may also be faceted though time and presented as an animation of the changing weather through the day.

Although charts can be used to make visually compelling additions to narrative, especially in the context of Big Data, the importance of supporting inferences with appropriate statistics cannot be overstated. Bivand and Krivoruchko (2018) offer examples of how inappropriate models, coupled with poor understanding of how Big Data has been collated, can lead to misrepresentations of the data. Their examples include misleading graphics of radiation levels around the Fukushima nuclear reactor, dissolved oxygen observations of the Pacific Ocean, and rainfall in South Africa. The key being that an understanding of how the data represents real world is essential to creating informative representations of underlying information, a point noted several times in this review.

No discussion about visualisation of data would be complete without reference to the work of Tufte (2001) which describes good practice, and failures, in the communication of technical information. While this reference has many examples of PowerPoint and Spreadsheet driven charts, it has much to offer in terms of Big Data summaries. The work of Edward Tuft has influenced all forms of visual presentation, and those pertaining to R have been collated by Piwek (2015) as a curated website with many examples. As an aside, the author produced this resource using R markdown and the `tufterhandout` package which provides the Tuft inspired output formatting used on the website. The Visual Literacy website of Eppler (2020) is also a rich resource on this topic, especially in relation to computer generated graphics that are likely to be created though the analysis of Big Data. Clearly, much intellectual capital has been invested in the process of visualisation and communication of knowledge derived from data and some specific implementations of tools to assist presentation are freely available for general use by informed users. However, their effective use is predicated upon a sound understanding of the data context and interpretation.

The relationship between types of visualisation and Shneiderman's data taxonomy supports a simplified generalisation of data verification within the interpretive process. This assertion

arises from accepting that real world data may always be mapped into the seven Shneiderman categories. Since each of these categories may be visualised using one or more of the charts listed in Tables 2.1 — 2.3, an understanding of data context provides the basis for verification. As an example of why visualisation is important, Brodie (2020), reported on the omission of positive COVID-19 cases due to 'exceeding maximum file size' in spreadsheet data. Taking the explanation at face value, a simple bar chart showing the number of cases would have shown numbers clipped to the same number, which should have raised questions before the issue became a problem.

A more complex example is taken here from Appendix F.2, which illustrates how real data may often comprise attributes that fall simultaneously into several categories, which corresponds to Shneiderman's manifold category. In such cases multiple visualisations may be used to explore a dataset and verify that the data appears to match the expectations from its real world context. In this example, the data are biodiversity observations from the NatureSpot charity, and interpretation of three plots in Figure 2.5 raise multiple interesting questions about the data. However, since the intent of this discussion is to convey the importance of high level visualisations as part of the analytical tookit, it will be concluded by noting that the elements in Figure 2.5 were derived from a single rectangular dataset comprising 249,322 rows, which makes manual checking strategies impractical. These observations influenced the practical implementation of templates described in Chapter 5 which describes how and where visualisations are used as a running verification check.

2.8 Most Significant Authors

During the reading of papers a small group of authors stood out as repeatedly contributing work in the domain of interest and influencing the direction of this research. Referring to Figure 2.6, the work of Crawford & Boyd led to understanding the need to always keep the real world context in sight and not to regard data as an isolated entity. This shift in viewpoint helps avoids the automatic assumptions about 'faulty' data impeding analysis, rather than exploring the possibility of limitations inside the analytical process. Wickham's work has proved an invaluable building block in the realisation of this research with the mapping of data into the ubiquitous rectangular format, so this work is shown on the right-hand side of Figure 2.2. While not a standard, it is the basis of creating locally consistent data

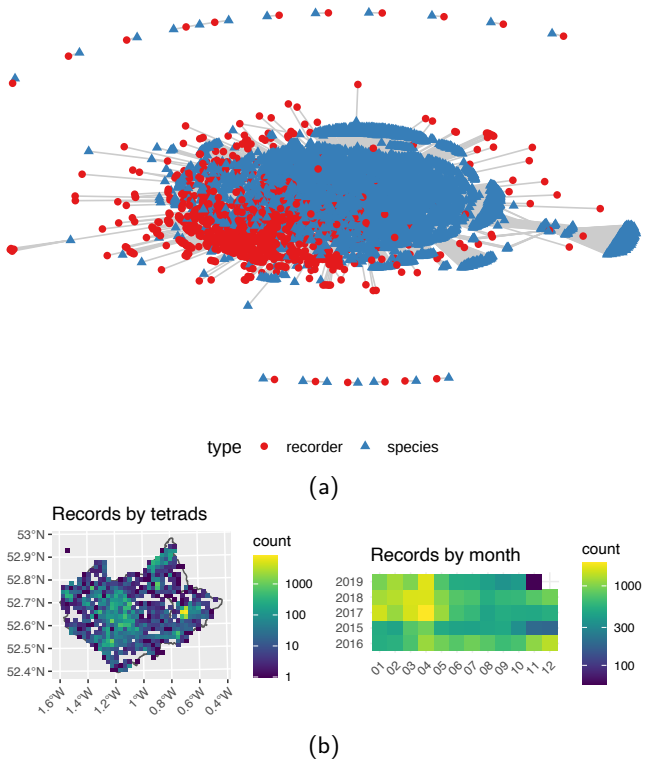


Figure 2.5: Three test visualisations of the same data illustrating the insights to be gained from high level overviews.
(a) A bipartite network view of biodiversity data with a Fruchterman-Reingold layout, uncovers relationships between observers and the species that they observe.
(b) The same data plotted with geospatial and temporal views offers further insights. The coverage is far from uniform with gaps within the county boundary and records outside it. The temporal distribution is limited and non-uniform too.

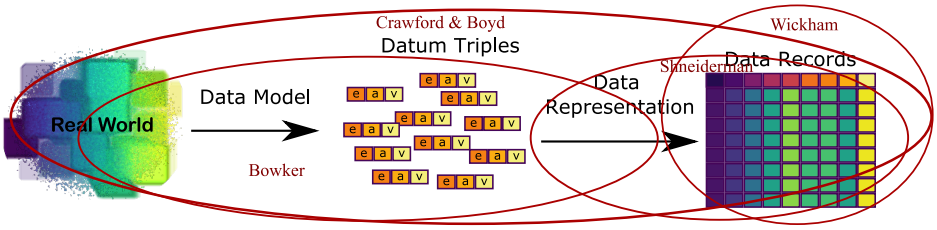


Figure 2.6: Mapping significant Big Data topic authors shows the overall coverage when overlaid on to the data definition diagram.

for manipulation and analysis. His implementation of a 'grammar of graphics' solution in ggplot2 enable functions that describe complex visualisations to be saved, and then consistently applied to rectangular data. This highly abstract approach using 'non-standard evaluation' (Wickham, 2019) is key to being able to effectively implement Shneiderman's ideas, as demonstrated in Figure 2.5. Bowker's work led to questioning the established presumptions of 'cleaning' 'dirty data' by reflection upon the overall data journey and ultimately to the need to develop the template theory described in Chapter 5.

The following short cameo on each of the authors provides additional background:

Kate Crawford & Danah Boyd Cited 2315 times, a paper written while both authors were base at Microsoft Research, explores issues beyond those that are purely technical, covering social and political aspects of Big Data (Boyd and Crawford, 2014). This thoughtful paper raises questions on the social impacts of Big Data, both positive and negative, while not providing answers, includes ethical aspects of the debate not included in solely technical papers. These authors have continued to contribute to this area research.

Hadley Wickham Cited 14587 times. As the lead architect of the Tidy R approach, an enabler of reproducibility, Wickham has made the R analytical language accessible to a new generation of data scientists because of the consistent syntax this approach has brought. He is a frequent contributor to on-line forums and his well-thought-out answers to questions undoubtedly have raised the profile of his work. His contribution to the 'ggplot grammar of graphics' complements the work of Ben Shneiderman.

Ben Shneiderman The author of many papers on Big Data and visualisation has been cited over 77960 times. His contribution is discussed in Section 2.7. One of the most striking aspects of his work was his anticipation of the challenges associated with Big Data, especially the need to retain the connection to the underlying detail, which couples perfectly with the work of Wickham.

Geoffery Bowker The book ' "Raw data" is an oxymoron' is one of the few works that seeks to explore the question: 'Where does data come from?' Although this book is currently cited less than 1000 times, it identifies a key gap in the data journey.

Returning once again to our definition of data in Figure 2.2 it is notable that Shneiderman's

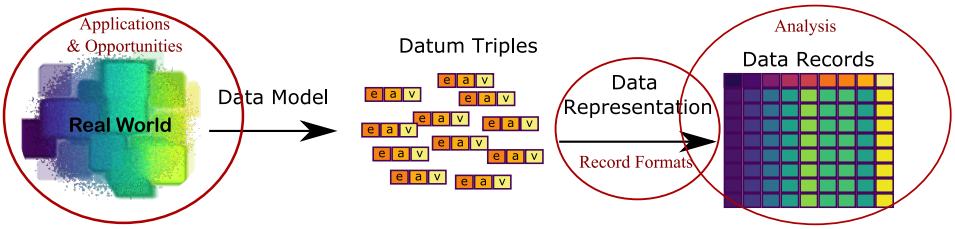


Figure 2.7: Mapping Big Data reviews on the data definition model reveals an incomplete coverage in the journey from real world observations to data records.

focus is on the visualisation and representation of data. Wickham's focus is data analysis; only Crawford and Boyd focus on impact to the real world, leaving little work, other than Bowker et al. (2013), that considers the overall connection between the intermediate stages.

2.8.1 Notable Reviews

Overlaying notable reviews onto the base data definition diagram in Figure 2.7 reveals a similar gap in focus that avoids discussing the Data Model and Datum Triples in the journey from real world observations to Data Records. The treatment of 'data cleaning' as a stand-alone topic as indicated in Figure 2.1 obfuscates the connection to the Real World effectively absorbing all other issues as 'Dirty Data'. This is not to say that there is not much useful work to be done in this area, but rather, a clear separation needs to be made between issues in the transformation process and those which are faults within the data. For example, multiple Data Models might be used to transform related observations using different Data Representations into Data Records. If the Data Models and Representations are perfectly described, and the transformations perfectly undertaken, the Data Records are in an inconvenient state, rather than faulty. This line of reasoning is developed further in Chapter 4.

The following cameos group Big Data reviews about Big Data into the categorisations overlaid on Figure 2.7:

Real World Applications Chen and Zhang (2014); Wu et al. (2014); Lewis et al. (2013); De Mauro et al. (2016b) agree that many scientific fields have already become highly data-driven with the development of computer sciences leading to a new paradigm of Data Intensive Scientific Discovery (DISD). For instance, astronomy, meteorology,

social computing , bioinformatics and computational biology are greatly based on data-intensive scientific discovery as large volume of data with various types generated or produced in these science fields.

Data Record Formats Chen and Zhang (2014) list technologies such as: Apache Hadoop; Apache Mahout; and Pentaho business analytics, in their list of technical challenges. However, these could all be considered more generically as implementations of Big Data compatible: infrastructure; machine learning; and business analytics tools and services. Wu et al. (2014) takes a more generic data centric approach and acknowledges the need to work with ' sparse, uncertain, and incomplete data' as a specific technical challenge. Lewis et al. (2013) go a step further and proposes a pragmatic hybrid approach using multiple tools and technologies to get required results. Only De Mauro et al. (2016b) applied a formal methodology to search through literature to name the challenges and skill sets that address them.

Data Record Analysis Dryden and Hodge (2018) note the challenges faced by the statistics community in response to big data, and Lazer et al. (2014) 'The parable of google flu: Traps in big data analysis' provide a telling example of how the analysis can go wrong. The collation of data in advance of defining the analysis also presents methodological challenges, and only the work on causal statistics developed over the past twenty years by Pearl (2020) seems to have the potential to introduce radical improvements by incorporating data context into statistical analysis.

Real World Opportunities Digital forensics is suggested by Guarino (2013) as an example of a new opportunity driven partly though the accumulation of data and partly by the growth of analytical techniques. The curation of data to forensic standards is an issue, particularly given the volume involved. This view is supported by Russom (2011) who saw advanced data visualisation and analytics as the biggest opportunity on the horizon. No suggestion is made as to the tools that might meet these opportunities, so perhaps the report should be interpreted as a business guide rather than a technical roadmap.

Typically, analytical tools are mentioned in passing in these reviews, and none have been found that focus on desktop analysis of Big Data as a desirable approach, a presumption challenged by the author in the opening paragraphs of this chapter. Indeed, Wu *et al* dismiss

without reference or justification, the use of PC based big data analysis with the following statement:

> …For small scale data mining tasks, a single desktop computer, which contains hard disk and CPU processors, is sufficient to fulfill the data mining goals. Indeed, many data mining algorithm are designed for this type of problem settings. For medium scale data mining tasks, data are typically large (and possibly distributed) and cannot be fit into the main memory..... *For Big Data mining, because data scale is far beyond the capacity that a single personal computer (PC) can handle*, a typical Big Data processing framework will rely on cluster computers with a high-performance computing platform Wu et al. (2014) …

This contrasts with already mentioned pragmatic approach of Dass *et al* from Statistics Netherlands, who evaluate Big Data as potential source for official statistics using a modest PC workstation and R (Daas et al., 2015). Their findings show typical Big Data issues including missing data, volatility and selectivity, which all need to be addressed before the analytical results can be accepted [7]. Overall they conclude that the official statistics community can greatly benefit from the possibilities offered by Big Data. The continuously improving capability of desktop computers due to 'Moore's Law' and related technologies have been well documented (Palmer et al., 1999), thus it is likely that they will be seen as increasing in utility with respect to Big Data (Manovich, 2011).

2.9 Questions

There are many challenges created by the data deluge that is Big Data. Technical aspects include: Data capture and storage, Data transmission, Data curation, Data analysis, Data visualisation (Chen and Zhang, 2014; Sivarajah et al., 2017). Societal challenges include: how to guarantee 'equity' among all citizens; manage fake news and opportunistic behaviours; balance individual privacy and general interest, accountability and responsibility with new opportunities (Azzone, 2018; Lewis et al., 2018).

The following paragraphs articulate gaps in the current state of the art around Big Data presented here as research questions. Due to the quantity and scale of the literature, many

[7] Note that the issues are framed as problems with data rather than of analytical process.

other important questions could be framed by focussing on each stage of the analytic process, or by considering more abstract societal and ethical questions, but those chosen are of relevance to desktop analysis of Big Data. Each question is derived from the preceding narrative, and is supported by a background description, along with a cross-reference to the sections of this report that guided the reasoning process.

How the analytical process be made reproducible and reusable? There are many challenges in achieving the goal of reproducibility: No analysis is likely to be a linear series of processes, each used only once. Analysis may be exploratory rather than directed, and data sources frequently updated. Re-usability may be desirable to repeat an analysis with a different subset of data from a primary source, such as a geographic region or time range, or even different data with similar characteristics. (See Sections 2.6 and 2.6.)

How can transformations be recorded and verified? All aspects of the data analysis process transform the data; for transparency and reproducibility every transformation needs to be recorded in a way that makes verification easy and alteration difficult. The source data also requires unique identification to support the requirement of reproducibility. This is not to doubt the honesty of actors, but rather a need to protect integrity in the analytical process. (See Sections 2.5, and 2.6.)

How can mischievous alteration of data be prevented? If data is shared, how can data users verify that the data has not been altered, from the intended original? Changes might be subtle, creating a bias not present in the original data. Provenance of code should be easy to verify, and freedom from malware a *de facto* expectation. (See Section 2.8.1, 2.5, and 2.6.)

2.10 The Elephant In The Room

There is a question that is so obvious that it is rarely alluded to, and seldom answered:

> What is raw data?

This is asked here in a pragmatic sense, rather than seeking a philosophical answer about the nature of data creation. The definition of data adopted in Section 2.2 and illustrated

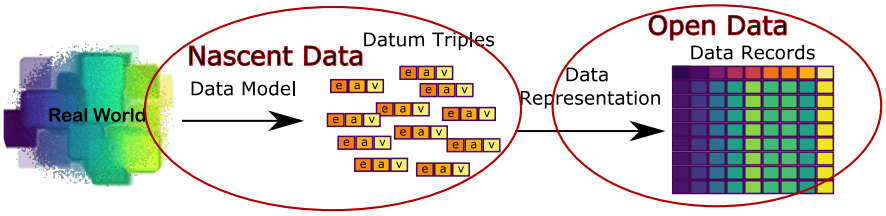


Figure 2.8: What is raw data and where does it come from? Seeing raw data as a combination of data model and datum triples places it outside the scope of all the big data definitions cited in this review.

in Figure 2.2 helps to understand what data is and how it connects to the real world, as collections of data triples: entity, attribute, value, and we can use this diagram to identify that part which is raw data as in Figure 2.8. Seeing raw data as a combination of data model and datum triples places it outside the scope of all the big data definitions cited in this review, making it an interesting concept to explore.

The list of research questions derived from the literature review avoid recognising the gap presented by this question about raw data, and the problem it represents by assuming an unqualified concept of 'data' as the starting point in the analytic journey and the problems that may be hidden as a result. A practical example is found within the review methodology in Appendix A. This noted how the two search engines used returned results as a series of files in different formats even though they contained essentially identical information. This is not an isolated problem, even though it has been ignored by most authors or dismissed as faulty data. For example, Daas et al. (2015) framed this type of problem as a data issue, but only Bowker et al. (2013) saw this as a gap in analytical process. The starting point for nearly all the academic work reviewed assumed that data were monolithic entities of data records and considered variance from this state as undesirable imperfections using terminology such as 'messy', 'untidy' or 'ragged' requiring 'cleaning' as preparation to become the *raw data* for analysis.

Thus, the fundamental gap identified by this literature review for this research is:

How can real world data be transformed for analysis? Recognising that most data exists in a primal disparate state is hidden in plain sight through the use of pejorative terminology presenting a presumption of *imperfect data* hindering analysis, rather than a fault with the analytical process. This is akin to the 1991 British Rail excuse for

> delays due to the: 'wrong sort of snow' rather than correctly ascribing the problem as being due to a lack of 'snow blowers' in South Eastern England which were needed to remove the unusually dry powdery snow that had fallen (Ayto and Crofton, 2009).

This leads in turn to the primary research question asked here: Is it possible to create tools that gather, transform and analyse raw nascent data that can be used without specialist programming skills? A secondary pragmatic question follows naturally from the first: Will data stakeholders use these tools?

Chapter 3

Methodology

Developing the underpinning philosophy for this methodology has been an iterative process. The following sections document the final standpoint and include the alternative choices and reasoning as a linear process; the key decisions are depicted in Figure 3.1. While this does not reflect the twists and turns in the decision-making journey, it is a simplification that makes the overall process much easier to communicate. Before moving on to the formal part of the narrative, this section starts by capturing the author's initial philosophical viewpoint which was strongly influenced by many years of engineering experience where the goal was always to: 'solve the problem', that was stopping the system from working.

Figure 3.1 is 'top and tailed' with two boxes referencing the EU funded ROAD2CPS project which inspired the questions leading to this research and why focussing on working within the constraints of available hardware and software tools yielded such good outcomes.

3.1 Initial Philosophical Viewpoint

An engineer's viewpoint is typically pragmatic in nature and strongly influenced by experience. Figure 3.2 incorporates a range of philosophical paradigms and includes choices within ontology, epistemology and axiology. Many other options are possible along with nuanced combinations that may be relevant to particular situations. Note that Pragmatism is a recognised paradigm (Saunders et al., 2016) and lends itself to mixed methods research.

Figure 3.1: Key decisions in the development of this research agenda and selection of a methodology.

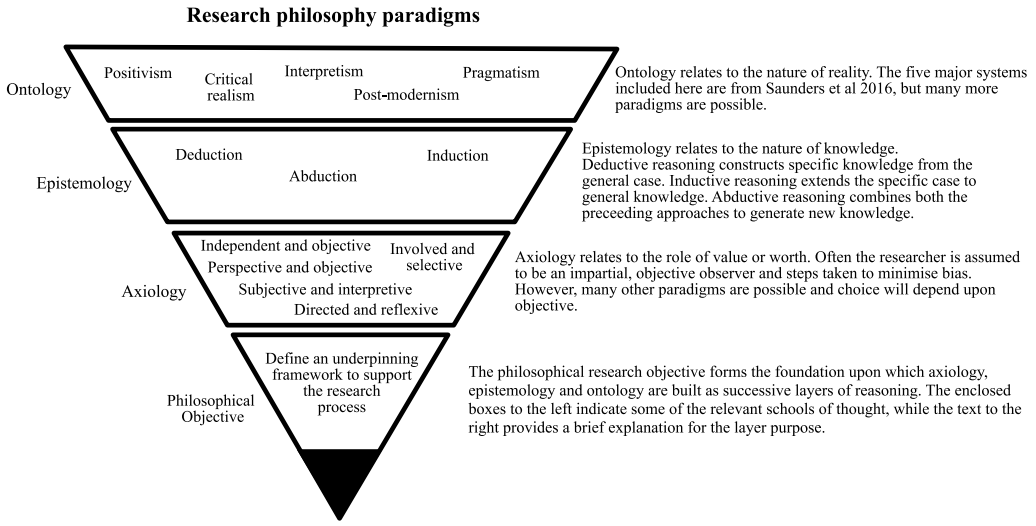


Figure 3.2: Philosophical paradigms.

As this work could easily be considered as mixed methods due to the wide variety of data sources and stakeholders, pragmatism was adopted as an initial viewpoint against which other options were tested, starting with an exploration of the implications of this viewpoint.

Pragmatism leads to an emphasis on finding 'solutions' rather than creating generalised knowledge that has relevance beyond the specific problem space, created by framing a research question as a problem to be solved. In this sense it tends to be deductive in nature, which in turn leads to the challenge of explaining why the solution contributes new knowledge beyond the confines of a specific question and the application of existing knowledge. While this approach may be relevant where a tightly defined research question has been framed, such that it may be addressed by well-designed experiments, it is not a good fit to more abstract questions without clear boundaries, such as those posited by the current work. From a philosophical viewpoint, this research is trying to develop a theory to understand the connection between nascent data of real world observations and data records used for analysis. The question being answered is: Is it possible to create tools that gather, transform and analyse nascent data that can be used without specialist programming skills? The relationship of this theory to the data definition introduced in Chapter 2 is shown in Figure 3.3. Note that the data model definition is external to this theory, although it provides the necessary function of mapping the real world into nascent data comprising entity, attribute,

value triples, which must be transformed for analysis. Expressed another way, the theory is constrained to apply only with data models that map the real world into nascent data comprising entity, attribute, value triples.

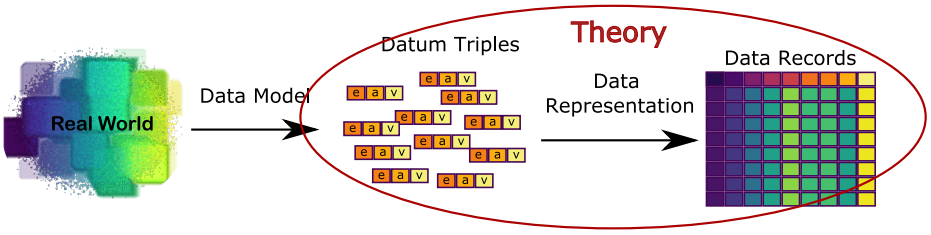


Figure 3.3: The intended coverage of the theoretical development is shown here mapped onto the data definition used in Chapter 2. The data model is presumed to map the real world into nascent data comprising entity, attribute, value triples, which must be transformed for analysis.

The following sections develop an understanding of the abstract nature of the research question and link it to a revised philosophical standpoint.

3.2 An Assumption Of Worth

One method to find possible alternative philosophical viewpoints is to ask: What abstraction uniquely defines the research area? The answer proposed here is to refine the domain of interest in terms of the **timing** of data collection, task definition, analysis and modelling. Expanding this construct suggests that this thesis focusses on cases where the collation of data are **prior to** task definition, rather than collected specifically for the purpose of experimental analysis, where procedures are likely to result in 'cleaner' data with fewer issues. A new term, *antecedent data*, is introduced here as a descriptor for the problem class of interest. The conceptual relationship of antecedent data to various analytical approaches is illustrated in Figure 3.4 as a vehicle explore the usefulness of this construct. Six types of analysis were proposed by Leek (2013) as archetypical analytical approaches. They are expanded here to help visualise compatibility of each approach with antecedent data and by extension, big data.

Mechanistic analysis techniques are well understood and supported by the statistical and mathematical approaches pioneered by Karl Pearson (see Plackett (1983)) and are often regarded as the 'gold standard' for scientific method, but may only be applied in carefully

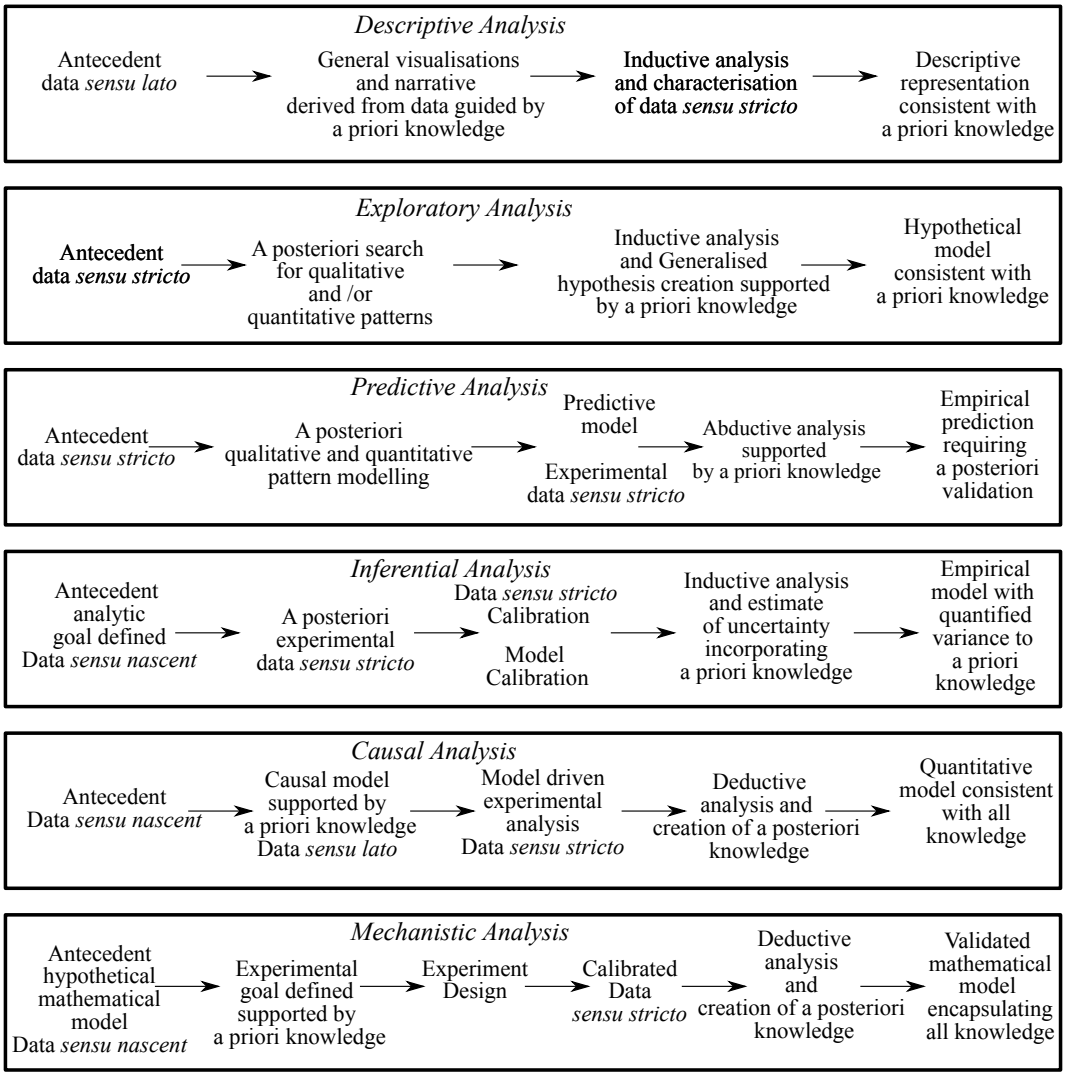


Figure 3.4: Analytic domain characteristics.

controlled situations where the data collation methodology are well understood. The causal analytical techniques developed by Pearl et al. (2016) are inherently compatible with the analysis of antecedent data and by extension, big data, as the models against which the data are tested, rely on *a priori* contextual knowledge for their definition. However, descriptive, exploratory and predictive analytical techniques are an equally valid part of the scientific method, and are inherently compatible with the analysis of antecedent data and by extension, big data, since these approaches do not imply constraints due to size, or any other parameter.

The author suggests here that using antecedent data in any process implies a belief that

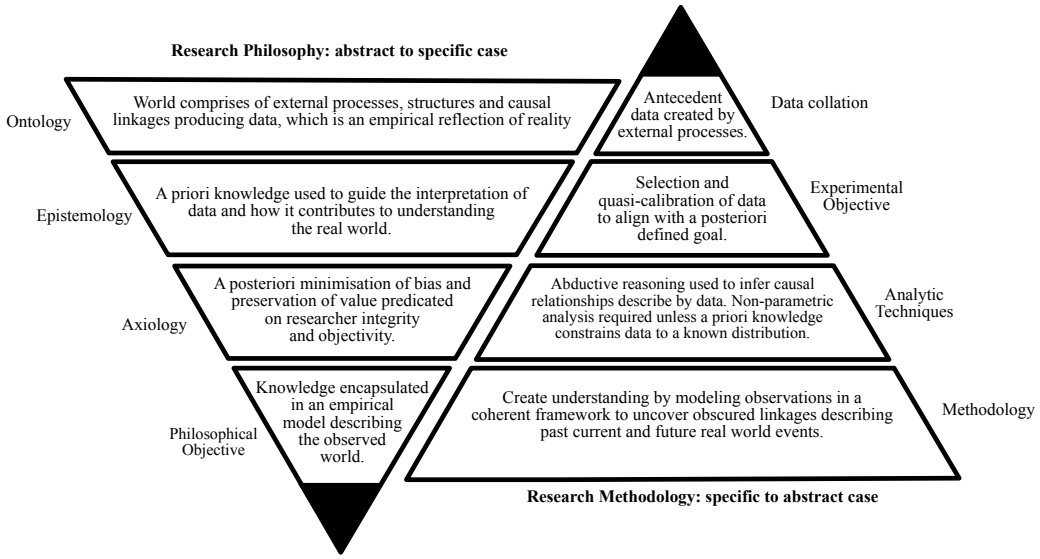


Figure 3.5: Relationship of research philosophy to methodology.

there is an extractable abstraction of worth in the form of knowledge contained within it. This is not surprising where data are collated as part of a designed experiment, but is it true for all the cases illustrated in figure 3.4? Accepting this assertion carries with it a belief that the world comprises external processes, structures and linkages which are empirically reflected as worth within data. This is not to say that worth is proportional to data size, or that it has an equivalence to monetary value, but instead provides the basis for a world view placing data in the overall context of human endeavour alongside other constructs with abstract worth, such as art. For example, social media data are frequently used as examples for big data analytical techniques; the implicit value in such data is the clear connection between the data and real world social interactions.

3.3 A Revised Philosophical Viewpoint

The preceding section introduced the concept of 'worth' as a foundation for justifying an analytical process, and this is now explored in more detail.

Data are an empirical representation of events and must be transformed and interpreted to provide a view of the real world. This type of philosophy, termed critical realism and

popularised by Bhaskar (2008), has an important role in providing an interpretive modifier between data and reality. Without this modifying philosophy, there is a danger that one could become trapped inside a world view based on false, misleading or mischievously altered data. It follows that understanding the chain of data transformations from source to the present is a necessary condition for protecting worth. The converse is also true: data may suggest a world view that is so different from that which is currently accepted as true, the new reality may incorrectly be rejected as worthless on the grounds that present data are not a true representation of the real world. A topical example is that of anthropomorphic climate change as reported by the IPCC [1]. Interpreting data and linking it to causal mechanisms may lead to an uncomfortable disconnect between the world in which we live and the world that may be a future reality without substantive and costly changes in human society. Thus, what constitutes acceptable proof of causality, and therefore valuable knowledge, may also depend upon personal viewpoint along with factors outside the experimental framework.

The assumption of worth contained within data also helps to extend the use of antecedent data into the lower two analytical domains illustrated in Figure 3.4: inferential; and causal analysis. In both cases data and mathematical models are used to perform an experiment with reproducible and interpreted results as part of an overall process in which knowledge is extracted.

Figure 3.6 translates these observations and assumption into a philosophical framework that strongly aligns to the critical realism paradigm. This revised philosophical paradigm, critical realism, is now used to develop a methodology that supports *a posteriori* techniques for knowledge extraction, rather than using *a priori* experimental design to test a hypothesis. The application of critical realism in supporting innovative theoretical constructs is described in the essay by Williams and Wynn (2018) which explores how to avoid becoming trapped by the current dominant accepted theory. The argument essence is that the conventional science script places too much emphasis on accepted wisdom and inhibits novelty. An alternative critical realism script allows the construction of a theoretical set of behaviours that describe observed events independently of other theories. Figure 3.5 illustrates the association between philosophical and methodological assumptions. These choices link the abstract world view to the specific objective of the research. Methodological choices link the desired outcomes to the abstract challenge of using data created by external processes outside

[1] Intergovernmental Panel for Climate Change `http://www.ipcc.ch/`

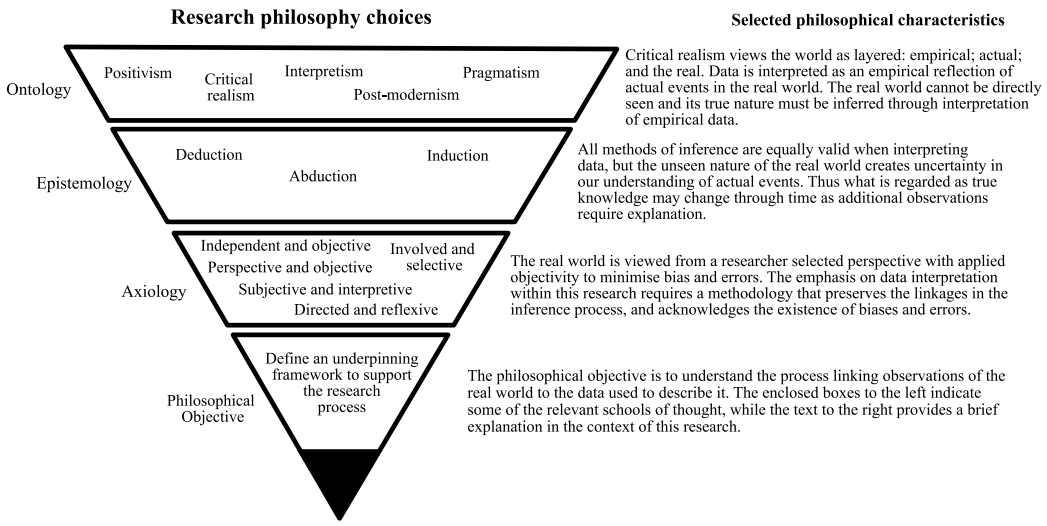


Figure 3.6: Philosophical choices made for this research.

of researcher control. The methodology details are developed in the following sections.

3.4 Applying The Research Philosophy

The concept of worth within data is used here to build a bridge between the adopted research philosophy of critical realism and this methodology, by making an assumption that the fundamental reason for all data analysis is to extract value and present it as knowledge. Thus, an actor will have objectives defined by a mix of personal and external drivers that may be met by extracting worth from the data. Interpretation will be guided by *a priori* knowledge and other factors that may be personal to that actor. The term 'actor' is interpreted here as a metaphor for an individual performing an analysis that recognises the variety of possible different, but equally valid, viewpoints for interpretation of results as knowledge.

Referring back to figure 3.6 *Philosophical choices made for this research.* will confirm that not only is critical realism consistent with the high level objectives of this research, but that it also provides a reference framework for interpretation throughout the research process. In particular, acknowledgement of the layered nature of the world is a reminder that uncovering the unseen and unknown can lead to changes in what are regarded as knowledge at any par-

ticular moment in time. This naturally leads to a layered approach for the research process that is presented here as three phases that reflect successive stages for the development of new knowledge:

Definition Building on the initial assumptions to develop a definition of the user requirements and an understanding of the current challenges;

Implementation Creation of tools that meet the user requirements and testing against assumptions;

Confirmation Evaluation of the match with user needs, and validation of the overall approach.

The review by Ekbia et al. (2015) highlights the many conceptual and practical dilemmas in the overall Big Data problem space, so for the purpose of research it is necessary to focus on a particular problem that has a general application. In this thesis, the straightforward device of a motivational example is used to keep the focus on solving the challenge, rather than the user interface in a suitable test domain. Social media has been mentioned several times in this work as a popular source of data with a defined connection to the real world, but while data are available through several providers for academic use, such routes lack obvious stakeholders. However, a direct plea for help in the biodiversity domain was made by Lewis et al. (2018) 'Wildlife biology, big data, and reproducible research':

> …There is also a pressing need for teaching computer programming skills that can expedite all of the above [analytic] processes and make them more reproducible… However, the vast majority of scientists are largely self-taught as computer programmers. We suggest that the most expedient means to incorporate these concepts into standard wildlife ecology practice is by incorporating them into standard ecology and wildlife biology educational curricula.…

Although Lewis et al. are based in Canada, informal enquires confirmed that local biodiversity data stakeholders recognised the same issues and were willing to discuss and share raw data.

3.5 Empirical Experiments

Local biodiversity data stakeholders always spoke as if there were a single database of records, this was not the case. Nascent data existed in many files and formats, while each record could be traced back to field observations, preparation for analysis was found to be a manually intensive process. Requesting access to data in the initial raw state was seen as surprising since stakeholders perceived the processed data, represented as 'Data Records' in Figure 3.7 as being more useable than the raw data for research actually sought. This perception of data value residing purely in data records of Figure 3.7 uncovers a potential bias in data offered for research in that there is a presumed requirement for 'clean' data by stakeholders that obfuscates the reality of the raw state. This can partly be explained by the concept of worth which was introduced in 3.2 to justify *why* an analysis might be undertaken, leaving the question of *how* open. Logically this leads to a duality of viewpoints when designing and reporting on empirical experiments in this research: answering a question, the why, with a successful analysis results in the extraction of useful knowledge that may be reportable in its own right from the view point of data owners. The how, that is, the techniques used to perform the analysis, may seem of less interest than the results, even though it is key to the goals of reproducible research and contain insights as to how the analysis may be extended to other domains.

Initial enquiries suggested that potential stakeholders were prepared to engage directly with the researcher keeping to a 1—2 hour time allocation, allowing structured interviews to be used as the method of gathering information relating to analytical tasks.

To guide the creation of a suitable motivational example, information elicited through interviews were used to guide the production of generalised problem statements, in turn leading to definition of requirements. It is posited here that ideal problems possess one or more of the following characteristics:

- Availability of large data;
- Usefully repeated with different data, or with subsets of data;
- Difficult to achieve with currently available tools;
- Useful to the stakeholders.

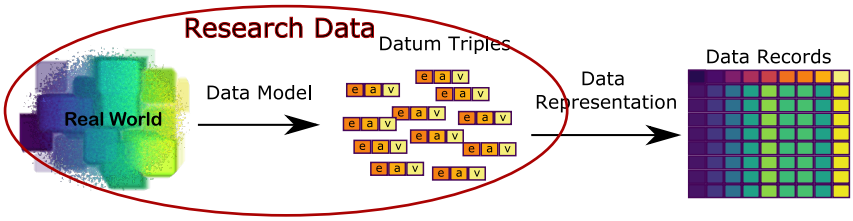


Figure 3.7: This conceptual visualisation of the required research data emphasises the need for data with a clear connection to the real world rather than data records with undefined manual interventions. Rather than design experiments to collate new data, this work will seek out untidy data that are already collected and difficult to analyse.

Conceptually these characteristics are represented in Figure 3.7 which illustrates the need to search for sources of untidy data that are already collated, rather than to design experiments that create new data.

The motivational example is developed in Chapter 4 that seeks capture the characteristics of a plausible range of end user tasks and associated data. A particular feature of this example is the narrative that describes how 'problems' within the data may arise due to internal and external influences outside the control of the data stakeholders. Using an example in this way shares some similarities with the well-established technique of using hypothetical personas for translating from the abstract to specific requirements (Billestrup et al., 2014; Almaliki et al., 2015) to guide the design of software interfaces. However, the aim of this research is not to develop a software application for an end user, it is to provide a solution to an initial step in the analytical process using software as a tool. The choice of a motivational example keeps the focus on solving the real world analytical challenge presented by the data, rather than as the development of user interface as a programming task.

This is represented diagrammatically in Figure 3.8 which illustrates how the stakeholders have a manually intensive task based approach to the descriptive analysis of the motivational example data. The research viewpoint uses a reusable template to take the data through multiple states and verification prior to analysis. This supports reproducible analysis because data *sensu stricto* are updated as additional data are added, allowing consistent approach for updating the output report. While offering interesting results to the stakeholders, the reproducible analysis is considered outside the scope of this research as the novelty lays within the concept of data state and templates. Issues relating to missing and duplicate data may be addressed using existing capabilities within the Tidy R approach of Wickham

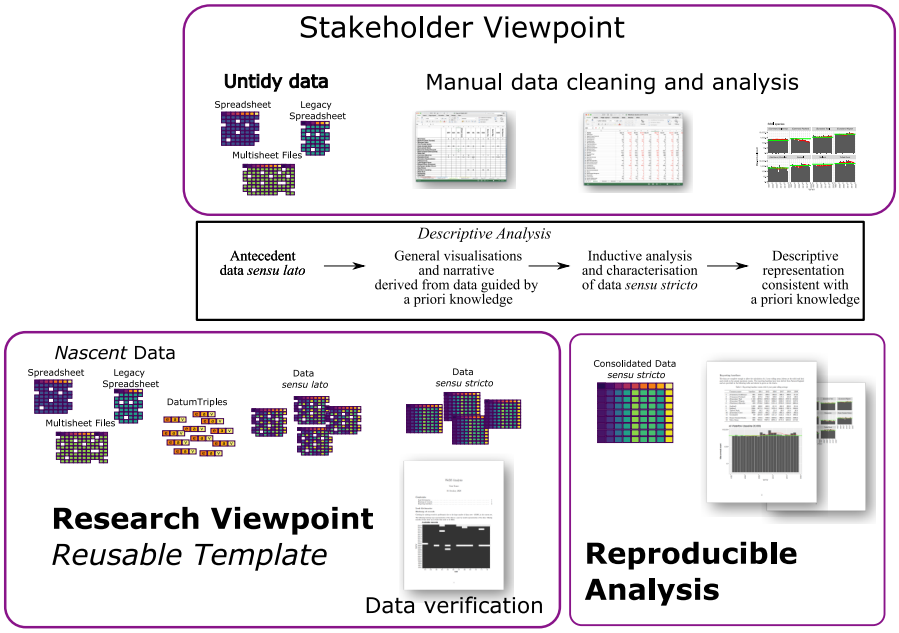


Figure 3.8: Stakeholders have a task orientated viewpoint based around manual cleaning and analysis of data. The research viewpoint considers the fundamental characteristics of data needed to realise the concept of reusable templates. The reproducible analysis that follows from the use of data *sensu stricto* falls outside the scope of this research, but it is undertaken to deliver results to the stakeholders.

(2014) and other analytical tools.

3.6 Implementation

Early work by the author indicated that the R analytical language coupled with 'analytical' templates were likely to prove an effective tool in the context of this research. The basic programming elements required to build such analytical templates were reviewed by Palmer et al. (2019) and found to exist, but no systematic implementations were known that addressed the issues with using nascent data, beyond mentioning 'data cleaning', the conceptual limitations of which have already been mentioned.

The initial *ad hoc* interaction with biodiversity stakeholders confirmed that the availability of data and subsequent analytic expectations have grown faster the capability to analyse the data. This research posits that less expertise is required to use templates than to write them. A reasonable approach to validation is therefore to determine if stakeholders will use

a template that meets their analytic goals.

3.7 Verification And Validation

As suggested in section 3.5, a duality exists between user goals and research objectives:

- Stakeholders are seeking to improve the analytical process and were noted in early interactions to readily accept and integrate third party analysis into their overall work. Volunteer contributions to support professionally directed tasks are a normal part of the working environment, so workplace procedures were already in place to support such help.
- Research objectives required the investigation of stakeholders willingness to adopt new tools and successfully apply them to analytical tasks quasi-independently.

An indirect verification process was proposed by working with stakeholders on current projects with a major analytical component and offering tools for adoption into existing workflows. A potential problem with this approach was that new methods would be seen as too much of leaning curve for adoption. There is therefore a risk that this work could successfully deliver useful and even publishable analytical results for the stakeholders, while failing to facilitate the adoption of reusable templates, and any long term improvements to analytical practice.

The proposed method to mitigate this risk was to work with the stakeholders on a regular basis accepting and resolving day-to-day analytical tasks, but using 'reproducible research' techniques for their delivery, while taking care to acknowledge the need to achieve outputs that can be integrated into exiting stakeholder workflows. This approach allows a deep insight into activities and motivations through formal and informal discussions. The record of these interactions will become the evidential basis for evaluation of research impacts.

The planned adoption of a reproducible template approach into new projects by stakeholders would be regarded as strong evidence for validation of research outputs. A request for ongoing support in analysing results would be regarded as weak evidence of validation, although it would be clear that the outputs were seen as useful. In this latter case, additional interpretation of the barriers for uptake would be ascertained by interviews.

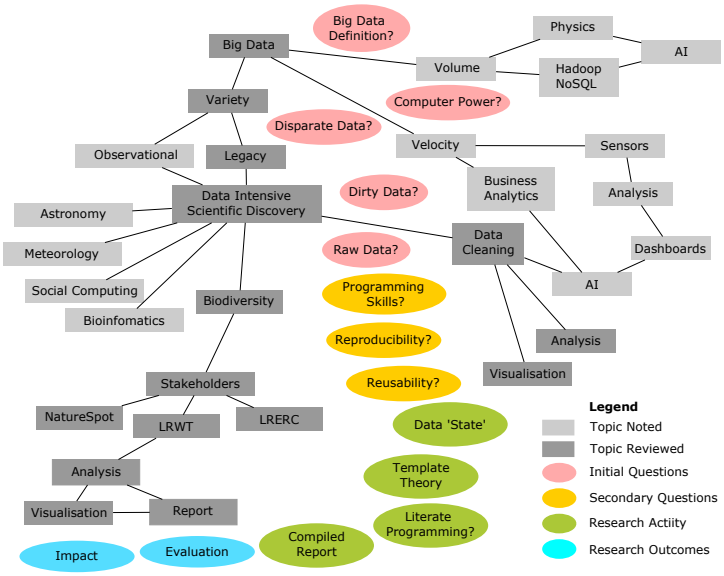


Figure 3.9: The contextual relationship of research activity and topics highlights why it was necessary to limit the scope of this research.

3.8 Methodology Summary

The key decisions made during this research were summarised in Figure 3.1, but it is helpful to revisit the topics uncovered in Figure 2.1 and overlay the research path as the series of 'stepping stones' in Figure 3.9. These serve to highlight that potential topics were not included simply because of the need to focus on a bounded research question, and not because of any perceived lack of merit. This also allowed for an element of curiosity driven selection of choices in research direction.

From the definition of a research question [2] choosing Creative Realism as a research philosophy is a natural choice because of the emphasis on real world observations. The research therefore requires access to data and stakeholders with analytical challenges to test the empirical theories that are constructed under the Creative Realism umbrella. A motivational example is used as a scenario on which to base an implementation tool inspired by actual data and stakeholder activities. Evaluation of research outcomes is based on generalising the motivational example to other stakeholder scenarios and commenting on extended applicability outside the primary domain of research.

[2] Can an empirical theory of the knowledge extraction process be developed that guides the creation of tools that gather, transform and analyse *nascent* data?

Chapter 4

Data Stakeholders

This research posits that data collected before the definition of a question to be answered are frequently encountered in the real world. This is termed here 'antecedent data' and is shown placed in analytical context within Figure 3.4. The arrow from data to the succeeding stages embodies an implied transformation into a format suitable for analysis, that was ignored for simplicity in this figure. This transformation is *not* just a cleaning of imperfect data, it is a *change of state* from an inconvenient form into a convenient form, while conserving existing information. This empirical research therefore requires access to antecedent data upon which to operate as a demonstration of the techniques developed.

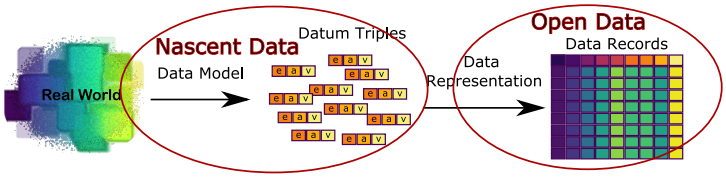


Figure 4.1: The nascent data required for this research falls within the envelope marked on the data definition diagram encompassing the Real World, Data Model and Datum Triples. Note that Open Data sources differ in that they contain processed data.

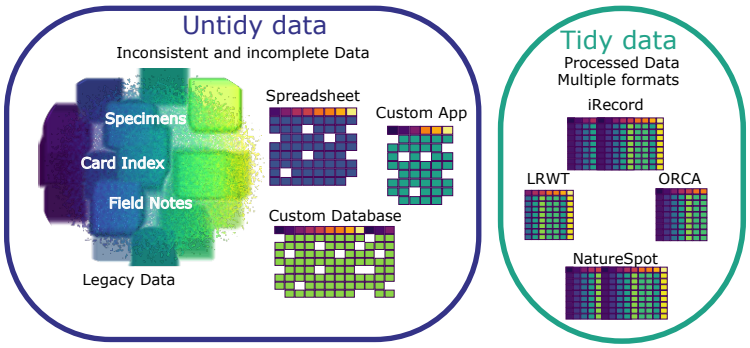


Figure 4.2: The biodiversity community holds large quantities of data in disparate formats which form the basis of evidential analysis. This 'Untidy Data' was incorporated into the methodology as the source of 'Nascent Data' for analysis.

4.1 Selecting Data For Research

The literature review highlighted the very broad range of situations which naturally lead to the creation of 'Big Data' leading to a broad range of potential stakeholders with associated analytical challenges. The gap identified in Section 2.10 relating to the processing of raw data becomes visible when stakeholders allow deep access to their data holdings along with all the associated shortcomings. To address this gap, access to original data in the messiest raw state was required but such access was not possible without the full support of the stakeholders. This requirement effectively eliminates many public domain big datasets from the list of potential candidate sources, although it should be remembered that such open data might be a useful complement to other sources.

To clarify this comment; access to open source large data such as post codes, company information, and geospatial maps is straightforward with clear licencing dictating usage. However, the expectation is that such data has already been processed and is in a 'tidy' well-defined state corresponding to data *sensu stricto* rather than *nascent* data, so is shown overlaid to the right-hand side on Figure 4.1.

Drawing on the observations made in Chapter 2 the following characteristics may be found in the 'nascent data' indicated in Figure 4.1:

- Too big to be handled in a straightforward manner.
- Messy underlying data in mixed files and formats.
- Contains multiple categories from Shneiderman's data taxonomy.

For the purposes of this research, rather than considering any of these characteristics as problems within the data, they are reframed as attributes of datum elements, which an effective the downstream analytical process must accommodate. Thus, this research requires data with a broad range of such attributes on which to verify the proposed theoretical approach.

These characteristic are frequently encountered within biodiversity data where longitudinal surveys may run for decades compounding problems with legacy data storage systems. The need for improvements in the data analysis techniques within the biodiversity community has been noted by Lewis et al. (2018) as a 'Big Data' and 'Reproducible Research' challenge for the Canadian biodiversity community. The arguments presented in this paper were informally noted to resonate with the UK biodiversity community indicating that they might be suitable data stakeholders for this research.

There are procedures in place for requesting biodiversity data from UK national databases, but this provides access to 'cleaned' records. A check using public credentials and the NBN Atlas [1] revealed that what is presented as a coherent national holding, is assembled from a huge array of local sources, each with its own stakeholders and access rules. Thus, accessing raw data would require negotiations with each of the stakeholders on an individual basis and while not impossible, this could potentially be time-consuming as the NBN would be acting as an intermediary. Instead, a direct approach was made to local stakeholders as every UK county has to provide mechanisms under 'The Natural Environment and Rural Communities Act' (UK Government, 2006) relating to the management and provision of biodiversity information.

[1] `https://nbnatlas.org/`

In Leicestershire, the home county for Loughborough University, the biodiversity community is principally represented by: LRERC, LRWT, NatureSpot, and Rutland Water Nature Reserve (Part of LRWT with its own staff). The diversity of data holding organisations from local government to charities matched the view formed by examining the NBN data. Each of these organisations were approached individually by the author, and each of principle stakeholders readily agreed to provide unrestricted access to data raw holdings. This included data that were regarded as too problematic for use in county level records and included access to staff so challenges and aspirations could be discussed. Figure 4.2 maps the data available from these stakeholders onto the conceptual requirements indicated in the Methodology in Figure 3.7.

The academic research question was recast into the terminology used by stakeholders as seeking to transform otherwise inaccessible data into a usable state. This clearly resonated as a real challenge as many hours of staff time were spent using manual processes to access such data, which meant that the author's involvement was seen as part of an ongoing contribution to stakeholder goals, rather than a demand on their time. The effort made in interactions with stakeholders to translate the academic goals of this research into the language used by the biodiversity community was benefical to all involved. In the end, this research was greatly enriched by the support of busy professional staff, and their recognition of the research question as a valid gap in their analytical processes.

4.2 Stakeholder Aspirations

Professional opportunities within the environmental and biodiversity community are seen as highly desirable and can attract many applicants. Selection from this pool of candidates leads to the appointment of highly talented individuals who are well-educated and excellent communicators, lucid both verbally and in writing. Volunteers may also have to go through a vetting process, which is especially stringent if contact with the public and children is part of the role, so both, professionals and volunteers tend to be well-educated and personable. These points are made because stakeholders may be either employed professionals or volunteers and no presumption was made about employment status in regard to role as a stakeholder for the purposes of this research.

Individual stakeholders for interview were identified through personal contact and secondary

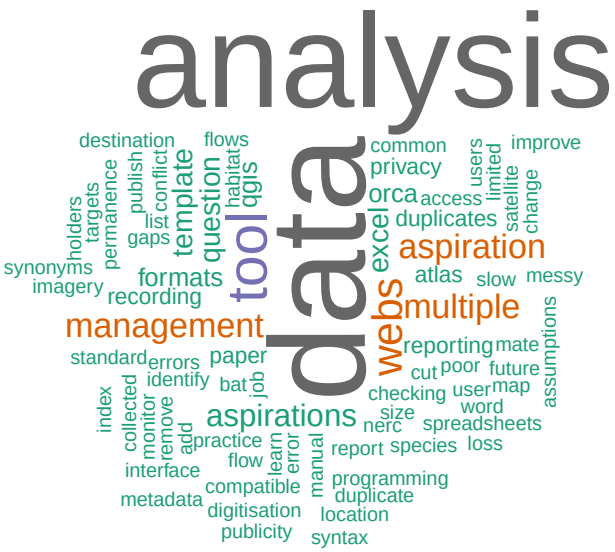


Figure 4.3: The word cloud of codes used to markup stakeholder interviews emphasises the importance of improving data analysis.

recommendations to identify suitable data wth which to implement the empirical experiments proposed in Section 3.5 of the Methodology. No presumption is made that stakeholders see themselves as data scientists or analysts, the focus of interviews were to identify analytical tasks that may be broadly classified within the definitions of Figure 3.4 *Analytic domain characteristics*.

This work is recorded in Appendix B which consists of the QDA annotated interview notes. The word-cloud summary in Figure 4.3 of the coding conveys a sense of the relative importance of improving data analysis. This is assisted by the analysis of bigrams (short phrases) presented as a network view in Figure 4.4 where the central importance of data is clearly highlighted. A few of the words need context to understand: WEBS wetland bird survey; orca county level biodiversity database; map mate recording software, but overall the sense is clear. Relationship with universities are also noticeable features, and interviews confirmed that these are seen as beneficial engagements.

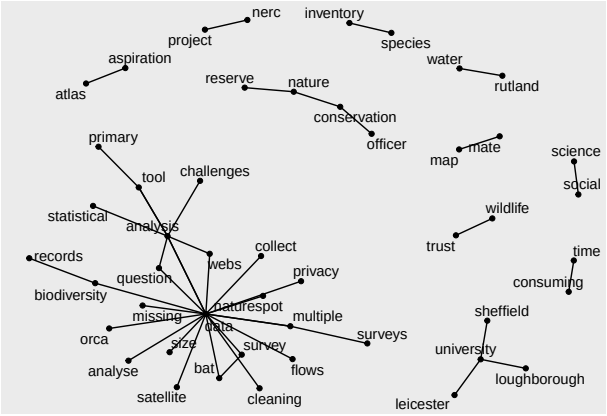


Figure 4.4: QDA analysis for bigrams, pairs of frequently used words, adds further insight to stakeholder responses. Note the range of words used in conjunction with data.

Key stakeholder goals were:

- Improve analytical techniques.
- Develop an understanding of the science behind data analysis of commissioned LRWT surveys.
- Provide evidence based interpretation of data to support environmental issues.
- Engage with the local research base.
- Support staff personal development.

These aspirations were a good fit to the research objectives and facilitated the development of a good working relationship.

4.3 The Use Of A Motivational Example

This section introduces the what, why, where and who associated with the example, but leaves open the details of analysis, which is described in a series of call-out boxes placed close to relevant text. This allows key mathematical and programmatic concepts to be supported with adjacent worked examples, rather than referencing material in the appendices. Techniques developed in the following chapters are explained using a worked example as

an illustrative device. It is helpful for this to cover a wide range of issues that might be encountered in any data, so while this example is inspired by real world observations, it is an artificially awkward set of data to work with.

The example is biodiversity themed, but the issues encountered are domain agnostic in that they are a consequence of the data collection over a lengthy time period, rather than any special attributes of biodiversity. It should be stressed that the issues with data are not intended to indicate inadequacies with the collection process, instead they illustrate how observations made over an extended period may encompass multiple enforced external changes to data structure and terminology. A formal scientific methodology is designed to prevent such issues over the data collection phase, but in the example scenario it is clearly not possible to repeat the observation process to provide a monolithic and clean dataset over the extended time period. As a generalisation, it is suggested here that similar problems to those in this example are likely to be encountered anywhere that data were collected in advance of analytical process design.

4.4 Example Background

This example is based on data comprising forty-years observational surveys of wetland birds species on a site in the Midlands of England in the UK. It is loosely based on the Wetlands Birds Survey (WeBS)[2] but the data are compromised to illustrate analytical techniques, so derived outcomes used as illustrations here *should not be used for environmental impact assessments*. Such caveats do not apply when the same analytical templates are used with valid data.[3]

A pseudocode description of the template associated with this example is presented in Figure 4.5 and includes marginal diagrams indicating idealised changes in data state performed by the template. The analysis of output Data *sensu stricto* is not core to this thesis, which concentrates on the issues of data preparation, but some analytical outputs are discussed in the example call-outs, and in Chapter 6. The theory underpinning template functionality is fully discussed in Chapter 5.

The volunteer surveyors monitored occurrences of non-breeding waterbirds since 1980. The

[2] `https://www.bto.org/our-science/projects/wetland-bird-survey`
[3] Valid source data are commercially available from the British Trust for Ornithology.

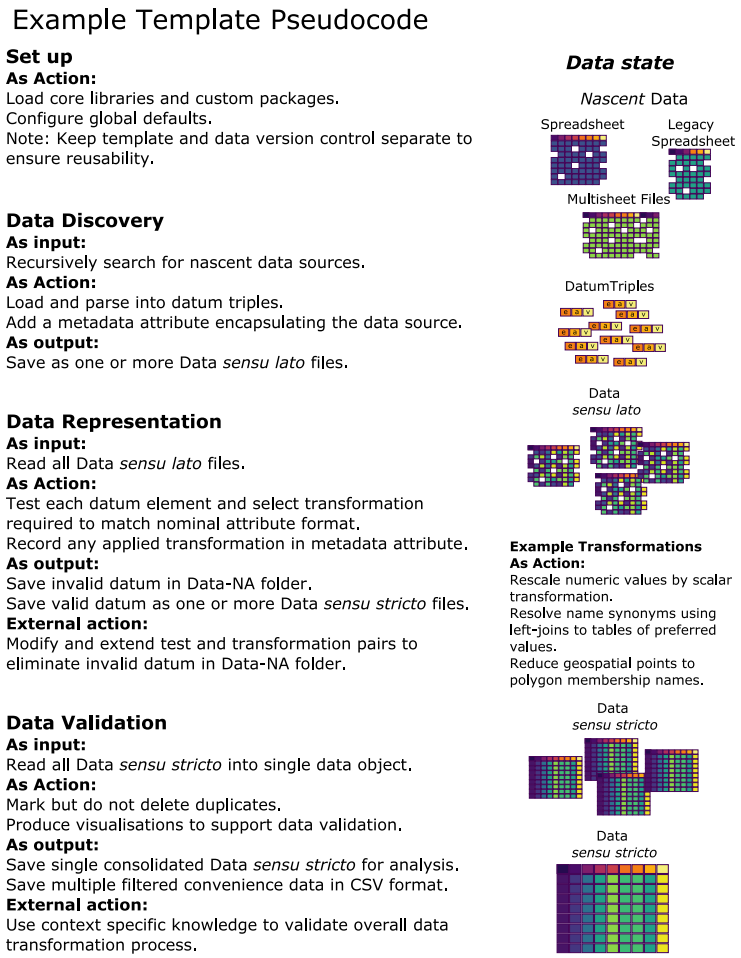


Figure 4.5: The motivational example template pseudocode includes a visual guide to the transformation of data state during code execution. When implemented in R markdown, each block of pseudocode corresponds to an executable 'code chunk' separated by context specific narrative.

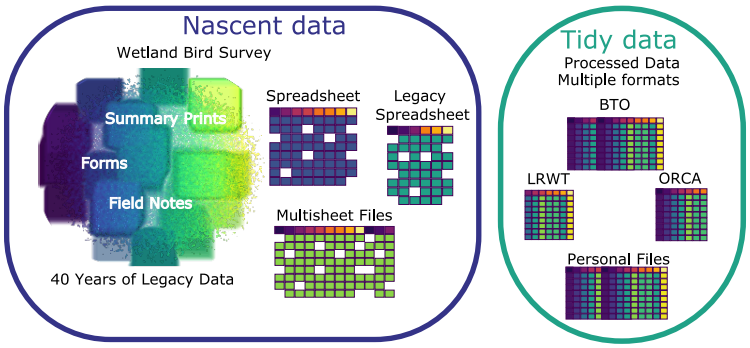


Figure 4.6: Conceptual visualisation of motivational example of nascent data which has accumulated many formats over time, and is inconsistent in format and incomplete in content.

methodology requires a monthly visit to a wetland site to count 'Waterbirds' a loose definition that includes wildfowl (ducks, geese and swans), waders, rails, divers, grebes, cormorants, herons and includes non-native, feral and vagrant species. During survey visits observation were made at specific locations and accurate counts made and recorded in field notebooks. Observations were transferred to a standardised form for collation. Older records were summarised by hand, more recently spreadsheets were used as part of the analytical process. While this may sound both straightforward and domain specific, the detailed description in the following sections uncovers hidden complexities and inconsistencies in the source material that are of the type mentioned by Bowker et al. (2013) as common issues with raw data. The starting place is termed here *Nascent Data* and is visualised in Figure 4.6 with a supporting description in Table 4.1.

The nominal data fields are those that form the primary intent of the data, without consideration of actual content. Generally, these fields are those that might be expected with this data, and it is the difficulty in formally pinning this data down that makes it nascent. Put another way, if the stakeholders were asked they would, with great confidence, state that the nominal fields in Table 4.1 were those contained within the data.

If data always conformed to the nominal state then the analytical process would be relatively straightforward and start with an extraction and cleaning type activity. However, this data were collated over forty-years, and while it is tempting to consider variances from the nominal state as issues to be overcome, instead we start with an empirical description of the *actual data* before moving on to designing a process to transform it into the *nominal format*, as

Nominal Field	**Nominal Content** ***Empirical Description***	**Actual Content**
Taxon	A list of species present. *Over the fifty-year period of observation both vernacular and scientific names have changed for some taxon. In addition, some taxon have been split, and others merged.*	This field comprises entries with many synonyms for single taxon and a date related interpretation of the precise meaning of other taxon.
Abundance	A numerical count. *While a numeric field is expected, sometimes notes are made commenting on the count.*	This field comprises numeric data interspersed with text comments.
Location	A name that cross references to a geospatial location. *Over time period boundaries and names have changed, and major habitat features evolved.*	This field comprises entries with many synonyms for single location so a date related interpretation is required. Geospatial references use multiple projections.
Date	Date entry. *Dates may be presented in several formats.*	There are multiple text and numeric representations of date that might be used.
Who	A text field recording observers as individuals or groups. *Interpretation of GDPR requires the omission of identifiable individual names from current and legacy records unless permission to include them is known to have been given.*	This text field may contain personal information that should be handled appropriately.
Notes	Notes are a miscellaneous long text field.	Text, but may contain mixed entries from multiple character sets and interfere with programmatic interpretation.

Table 4.1: Nascent Data. Note that while the intended information is contained within the six data fields that capture attributes, there are considerable differences between nominal and actual field content in this nascent data.

this is what the stakeholders expect, and finally present the data for analysis. This work diverges from other 'data cleaning' approaches by expanding on the concept of data *state* as a mechanism to guide the overall process.

An examination of Table 4.1 will show that the differences between nominal and actual content are complex and transforming the **actual** field contents into the required **nominal** field form requires sophisticated functionality. A strategy that attempts to read data using the presumed nominal fields will have many issues and consider the source data 'dirty', which is why the actual field contents are a better starting point for transforming nascent data. This subtle viewpoint shift is in keeping with the critical realism philosophy adopted by this work leading to an acceptance of the empirical data as a reflection of the real world.

The physical data format is neither consistent, continuous or monolithic due to the longevity of recording. At any particular time, the best practises were followed, but these changed through time. Table 4.2 lists the physical formats used encode the data which is spread in tens of files, some of which contain multiple work sheets each containing a single day of observations in a tabular format.

Format	Description
Field Notes	Handwritten source material transcribed into digital format.
<file>.csv	Comma Separated Value text file that may be created and read by many software programs. Other values such as <TAB> or fixed spaces may be used. Several character coding schemes may be encountered. ASCII which is a subset of UTF-8 is the *de facto* standard, but UNICODE (UTF-16) may be used in some Microsoft files.
<file>.xls <file>.xlsx	Data presented as spreadsheets, and multi sheet spreadsheet files in multiple Microsoft proprietary formats.

Table 4.2: Physical Data Format. Note how data *nascent* may be spread across multiple physical formats.

This motivational example serves to illustrate the complexities involved in analysing data that are superficially simple in content and structure. Although artificial, it is closely based on reality, and serves as a foundation upon which to develop techniques for working with nascent data. The following chapters will return to this example to demonstrate analytic approaches that start with the actual data state as presented in Tables 4.1 and 4.2 to create reproducible transformations that facilitate onward analysis.

4.5 Generalisation To Other Scenarios

The example is a wetland bird survey, and it is the *a priori* knowledge associated with this context that ensures transformation of state preserves information and the enables subsequent meaningful interpretation of the motivational example data. This assertion may be tested in a thought experiment by attempting to interpret the 'Actual Content' column of Table 4.1 in isolation from the context provided by the field name or nominal content description. Every imaginable route to interpretation requires an abductive process and external knowledge to build a context. The types of issue found in the data within this motivational example are on one hand context specific in detailed understanding, but universal in that every context will have a similar set of issues. Figure 4.6 attempts to illustrate how nascent data are almost always presented in many formats, not just neat rectangular arrays. This is why the subsequent chapters of this thesis introduce a novel terminology of data *state* to facilitate the generalisation of novel transformation and analytical processes. It is not proposed here that these states are reflections of the real world, but they are useful descriptive artificial devices that can be used to clarify discussion of a complex topic.

4.6 Data Stakeholders Summary

The empirical nature of this research required access to nascent data coupled with an understanding of its real world context, plus stakeholders to provide an understanding of the analytic challenges that may loosely be grouped under the heading Data Intensive Scientific Discovery, as shown in Figure 2.1. The biodiversity community were identified as potentially suitable from the literature review and later confirmed as a source. Interviews were used to develop and understanding of stakeholder goals.

A motivational example was introduced as a vehicle for research. This effectively couples the data to a real world context as required by the methodology, and is also used to help explain the shift away from the presumption of 'dirty data' requiring cleaning as a universal initial step. While the example relates to biodiversity, the type of issues with the data are unrelated to domain, even though a contextual understanding may be needed to explain the issue. For example, a geospatial location might be expressed as a place name, rather than co-ordinates.

Chapter 5

Template Theory & Implementation

The analytic goals have been introduced in Chapter 4 by describing stakeholder intentions, however, this uncovered gaps in the vocabulary needed to describe all the characteristics of data involved in the analytical problems that need to be solved, which in turn make it difficult for users to articulate a requirement specification. To address this problem, a theoretical understanding of templates is built in this chapter by considering the essential required characteristics for reproducibility and reusability following the structure in Table 5.1.

The theory is developed as a series of diagrams that are thought experiments that describe an idealised solution. Justification for producing a theory by constructing abstract structure in isolation from existing solutions may be found in critical realism essays by Williams and Wynn (2018) and others. The argument essence is that the conventional science script places too much emphasis on accepted wisdom and inhibits novelty. An alternative critical realism script allows the construction of a theoretical set of behaviours that describe observed events independently of other theories, as here. This has the advantage of not being bound to any existing platform, however, to be useful, we must still demonstrate that these behaviours can be grounded into a feasible solution. This is achieved here by using the motivational example for verification of actions that arise from the constructed theory. Final validation is undertaken by assessing how well the final results meet stakeholder requirements.

Section	Description	Maths	Example
Raw Data / Nascent Data	Nascent data is 'real' raw data and forms the starting point for the analytical process.	The mathematical representation coerces the nascent data into a rectangular form that is conducive to onward analysis.	Disparate data arises as a consequence of long term observations, from changing methodology and external events.
Data *sensu lato*	Data are filtered and assigned to nominal fields forming a locally defined group of similar data arrays.	The matrix representation of data *s.l.* is used to demonstrate useful properties of this state. Actions may be applied across multiple files including recombination and splitting.	The data may be presented as an untidy set of files with duplicate and missing data which may be too large to review manually. Bulk actions are possible that help guide the ongoing transformation process. Data that cannot be transformed is marked 'data-NA' for inspection.
Data *sensu stricto*	Data are transformed and filtered to strictly meet the nominal field definition.	The mathematical representation demonstrates how data *s.l.* may be transformed by the sequential application of transformations into the desired data *s.s.* state. Critically, additional transformations may be added without disrupting those that have already been applied.	Not all the required transformations are known at the outset. Ensuring the software implementation follows the iterative mapping of the mathematical description allows transformations to be developed independently of each other.
Implementation	Relating the mathematics to figures pulls theory into practicality.	The equations may be interpreted descriptions of functional code blocks.	A descriptive analysis is required to relate data *s.s.* to the context along with a method to assess the completeness of transformation.
A Practical Template Using R	Implied requirement for a user interface debugging capability. These have both been provided through existing RStudio and Markdown tools which are designed to support literate programming techniques.	Functional named code 'chunks' implement the mathematical elements alongside descriptive narrative.	The final report is woven from the data *s.s.*, code, and narrative. Adding more data will only require the updating of the narrative, not the code. We have demonstrated a template that may be reused without programmatic skills.

Table 5.1: Template theory chapter structure

The motivational example is used as an explanatory device to provide supporting narrative in a series of call-out boxes. The contents of each example box relate to the concepts discussed, and serve as link between underpinning theory and empirical observation of the real world. Example 5.1, introduces the motivational example as a messy data type problem that has issues as a consequence of long term collation. Subsequent call-out boxes illustrate problems and solutions that are addressed using templates based on the theory developed here.

5.1 Nascent Data

***nominal* data fields** This term is used to describe the **presumed** data fields and structure.

data *nascent* This is the **actual** initial state of data. Variances from the presumed nominal state are often described with pejorative terms such as 'messy' and 'untidy'.

data *sensu lato* Once transformed into a readable rectangular state, this raw data is termed as data *s.l.* to emphasise that data may need 'cleaning' or other transformation before use. Multiple instances of this state may be combined row wise to form a larger data *s.l.* set.

data *sensu stricto* Once data are transformed into a fully defined state ready for analysis it is termed data *s.s.* . Multiple instances of this state may be combined row wise to form a larger data *s.s.* set. However, if any data *s.l.* are included in such a combination, the result are data *s.l.*.

Figure 5.1: A lexicon of data states. Note that changing the state does not create or destroy information.

The term 'Nascent data' is used here to distinguish it from the term 'raw data', which is frequently applied to data which has already been filtered and transformed in some way, rather than to data which is truly in its 'raw' state. Justification for this viewpoint is found in work by Bowker et al. (2013) who collate a series of essays challenging presumptions on the nature of raw data. These illustrate scenarios across multiple scientific domains where 'raw data' has been transformed and filtered by processes applied prior to incorporation into a formal methodology. This should not be seen as an attack on the *accuracy* of data used in the scientific method, but instead is looking past the attractive simplicity of idealised data into the underlying real world complexity, and the need to describe *all* the processes needed to access that data. Figure 5.1 introduces the essential terminology in the construction of template theory. With this introduction in mind, nascent data are conceptually represented in Figure 5.2 as the starting point for the application of the template approach developed in this chapter. Example 5.1 provides the context for the data.

Transformation of nascent data into a form that may be used in analysis requires the application of *a priori* knowledge about the data context. The foundations for this assertion are

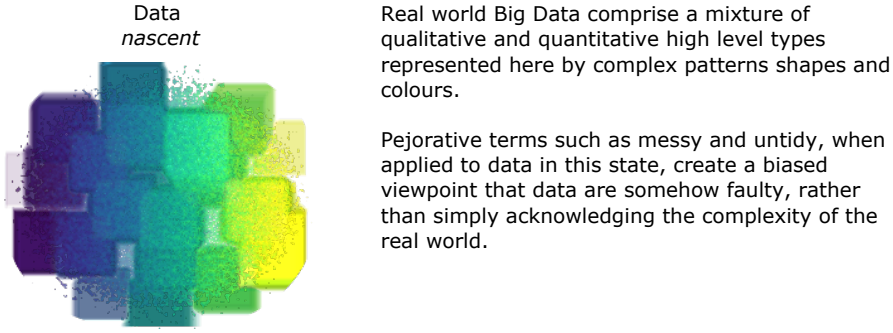


Figure 5.2: Introducing nascent data

made in by Kitchin and Lauriault (2015) who observe the importance of external knowledge to understand all the factors that shape data collation. This also emphasises why a story like detailed introduction to the motivational example in Section 4.4 is needed to apply deductive reasoning to inform the transformations required to use the data. Without a background narrative, we would only see nascent data as 'untidy data', and have to apply inductive reasoning to build a model that may be used to inform the transformation process. Given that practical model verification will require knowledge external to the data, we arrive back to the opening statement of this paragraph: *a priori* knowledge about the data context is required to transform nascent data for analysis. [1]

A useful pragmatic model, based on the work of Dadzie et al. (2009), is used here to describe the application of *a priori* knowledge. This is represented here as a 'Knowledge Filter' that is used to impose a rectangular matrix format on to the data based on this external understanding. If the question of *how* this might be achieved is set aside for the moment, the implications of what might be possible *if* a 'Knowledge Filter' were successfully applied may now be explored.

The rectangular matrix format for data is functionally equivalent to programmatic data array conventions for manipulating data, supporting the design of software to manipulate the data. The effective starting point for the analytical process developed here is anything that may be mapped into a disordered rectangular format, of the general form described by equation 5.1, without the need to make assumptions about the collective meaning of

[1] Knowledge rather than information is used to distinguish the reasoning used here from machine learning techniques which may be able to create algorithmic models from the same data. Whether such models can replicate the nuanced factors shaping the data that are understood from an appreciation of the data context is an open question.

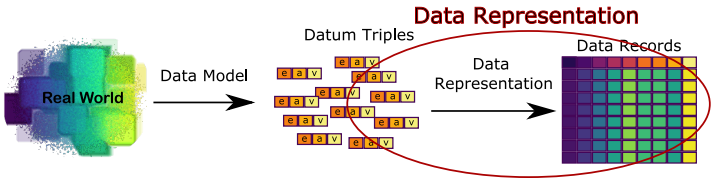


Figure 5.3: Domain specific assumptions using *a priori* knowledge are used to guide the representation of the real world data into a rectangular format.

the columns and rows. In this disordered representation, there is no association of position within the matrix with data attribute. This representation is closely related to the data definition of Figure 2.2 where the Data model is equivalent to the Knowledge Filter and the datum triples by $\mathscr{D}_{Nascent}$. Spreadsheets are a common example of digital data that may be presented in this format.

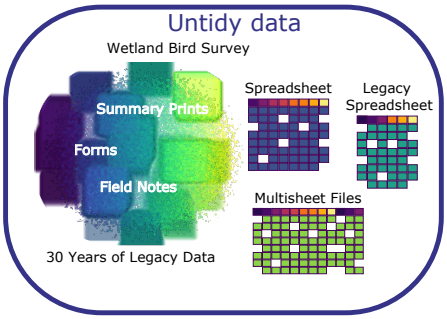
$$\mathscr{D}_{Nascent} = \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m,1} & d_{m,2} & \cdots & d_{m,n} \end{pmatrix} \tag{5.1}$$

To progress with the analysis we first reorganise $\mathscr{D}_{Nascent}$ into columns of common data types *sensu lato*. *Sensu lato* is used as a qualifier as the disordered data may use multiple formats and units to encode the same data type. Importantly, this qualifier allows us to usefully group data that we know to be related using external knowledge, without worrying about uniformity of representation. Typically, transforming $\mathscr{D}_{Nascent}$ into $\mathscr{A}_{Sensu\ lato}$ of equation 5.2 is achieved by repositioning elements into the desired arrangement. Thus, we can see that the purpose of the data representation in Figure 5.3 is one of identification and grouping of the data elements, rather than transformation of element contents. While mathematically trivial, we will show that this re-ordering imparts properties that simplify the onward transformation process and facilitates the construction of templates.

5.2 Data *Sensu Lato*

We now define data *sensu lato* as a set of one or more matrices of the form equation 5.2:

Using the motivational example introduced in Section 4.3, there is a need to develop action plans based on the interpretation of results from a formal monthly survey of a Wetland Bird Survey (WeBS) over three decades. As a consequence of the longevity of the survey, multiple volunteers have contributed to the effort, and collation spreadsheets are in a variety of Excel versions and internal formats. As habitats have developed, natural boundaries have changed, along with the names of units. Partly due to the success of habitat management, and possibly due to climate change, the species present have also changed.



That data are presented in multiple files and formats, containing multiple taxonomy definitions and synonyms that must be regarded as a feature to be addressed by the template approach as this condition is an empirical consequence of the collation methodology, rather than errors of process. In this example, the raw data may be a messy mix of duplicated entries, summary reports and other loosely related documents collated over time.

Example 5.1: The source of real data may be across many files and formats.

$$\mathscr{A}_{Sensu\ \ lato} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \tag{5.2}$$

Where the matrix elements $a_{m,n}$ represent data elements grouped as rows of observations and columns of similar data types.

The preceding section emphasised the need to reshape nascent data into a rectangular format represented as a matrix along with a future promise to justify the assertion that this will simplify the analytical process. The properties of data *s.l.* that arise from this representation are now explored and partnered with a conceptual template to show how reusability and reproducibility are supported within this theoretical construct.

Note that general form of equation 5.2 makes no assumption about the size of the matrix, but when interpreted using the terminology of Figure 5.1 it should be clear that the columns represent *nominal data fields* so multiple matrices of related data *s.l.* will have identical columns. It makes no difference to the information contained in the data if it is represented as a single large or multiple small matrices. Thus, data *s.l.* can be divided or combined row wise if they share nominal data fields, so even though the definition is very loose, hence use of the term *sensu lato*, this representation imparts properties to the data that were not present in the source nascent data. These properties of data *s.l.* are now examined in conjunction with reusable template actions.

5.3 Data *Sensu Stricto*

Continuing with the same style of notation, data *sensu stricto* may be defined as a matrix of the form in equation 5.3. As with data *sensu lato* the rows equate to observations, but now the nominal columns have all been transformed to self-consistent formats equating to 'tidy data'. Implicitly we are making an assumption that we can transform the data *sensu lato* (equation 5.2) into data *sensu stricto* (equation 5.3), an assumption that will be justified after the characteristics of data *sensu stricto* have been more fully explored.

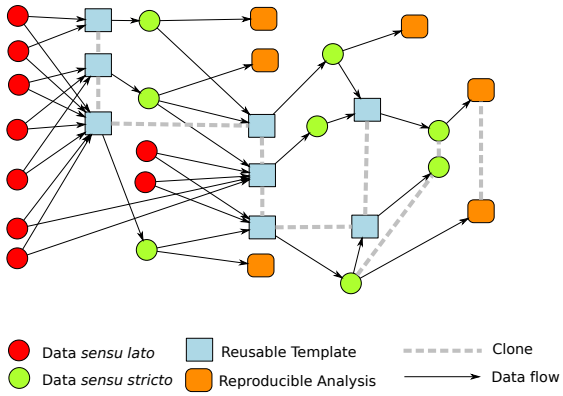
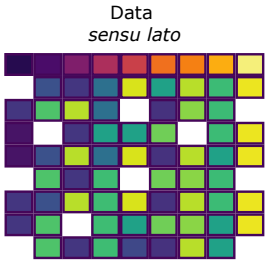


Figure 5.4: Essential characteristics of reusable templates in a reproducible context. Raw data *sensu lato* is transformed into clean data *sensu stricto* for reproducible analysis.

The data stakeholders are confident that the data from field notes and forms had already been transcribed into a spreadsheet format for reporting purposes. Over three decades software has changed along with the staff responsible for data curation. Although there is no consistency of layout, format or coding, the six nominal data fields of Table 4.1 may be grouped into columns to effectively transform the nascent data into data *s.l.*



Although this is an untidy data format with many imperfections that cannot yet be used in analysis, the process effectiveness is indicated by the number of data rows and undefined elements. Data context should provide an expectation of the number of rows, and undefined elements equate to unsuccessful data assignment into groups. These numbers are of no direct interest to the data stakeholders, but they provide a quantifiable guide to the programmatic development of the template designer, and a measure of its basic effectiveness in general use. For this reason data that cannot be transformed is marked 'data-NA' for inspection and deductive reasoning used to create a resolution.

Example 5.2: Data sensu lato.

$$\mathscr{B}_{Sensu\ \ stricto} = \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,q} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ b_{p,1} & b_{p,2} & \cdots & b_{p,q} \end{pmatrix} \text{ where } m = p \tag{5.3}$$

The matrix elements $b_{p,q}$ represent data elements grouped as transformed rows of observations and columns of common data types *sensu stricto*. That is, for the same data, the total number of rows of data *sensu lato* m are the same as the number of rows in the related data *sensu stricto* p. While we can define the equivalence in the number of rows as $m = p$, there is no basis on which to infer that $q = n$. We can explore why the number of columns in equations 5.2 and 5.3 might differ using the data context from the motivational example to assist the explanation. Dates are encoded in one of the nominal fields, and while this could be a single column with an ISO 8601 number string, other valid single and multicolumn formats may be used for convenience: year, day and week numbers might be encoded in separate columns. Other cases may also arise along with derived fields included to facilitate analysis but generally we would expect that $q \geq n$.

In the preceding section, we state that there may be multiple matrices of the form given by equation (5.2) to cover the complete data. The same reasoning may be applied to data *s. s.*, and since each $\mathscr{B}_{Sensu\ \ stricto}$ has identical columns, they may also be trivially combined row-wise.

$$s = 1 \geq \mathscr{T}_{Test}(a_{m,n}) \leq r \text{ selects from } \mathscr{T}_{Trans} = \begin{pmatrix} 1 \\ \vdots \\ r \end{pmatrix} \tag{5.4}$$

Where $\mathscr{T}_{Test}(a_{m,n})$ is a test for data type returning an index s, and $\mathscr{T}_{Trans}$ is a array of transformation functions.

We now introduce a function to describe the transformation of $\mathscr{A}_{Sensu\ \ lato}$ into $\mathscr{B}_{Sensu\ \ stricto}$ using mathematical representations. These equations are not intended to be externally justified, but instead should be interpreted as self consistent compact descriptions of a novel process. Equation (5.4) introduces a test that provides a single value used as an index to

select a transformation function from an array of possible functions.

In the following equations a $\bullet$ operator is used to 'pipe' outputs from one function to the next in a chain. The purpose of using this abstraction is to simplify the description of nested functions in Equation 5.5 from which Equation 5.6 is derived. This operator is implemented in programming languages, including R, so this representation is moving towards a form that supports implementation. Once we have tested the element and have determined an index for the transformation function, the transformation process for each element may be represented in the form of Equation 5.6.

$$a_{m,n} \bullet \mathscr{T}_{Test}(a_{m,n}) \bullet \mathscr{T}_{Trans}(s) \equiv \mathscr{T}_{Trans}(a_{m,n}, \mathscr{T}_{Test}(a_{m,n})) \tag{5.5}$$

$$a_{m,n} \bullet \mathscr{T}_{Test} \bullet \mathscr{T}_{Trans} = b_{m,p} \tag{5.6}$$

But what happens if there is no transformation selected by $\mathscr{T}_{Test}(a_{m,n})$ in Equation 5.4? Equation 5.6 is still valid, but the transformation does not exist, so we can say that the *value* of $b_{m,p}$ is undefined. This important definition will be used again after we iterate over the entire matrix and to apply the transformation of Equation 5.5 to every element. As there is no universal symbol for iteration, we define one here:

$$\forall x; x \in (1, \cdots, n) \equiv \prod_{1}^{n} \tag{5.7}$$

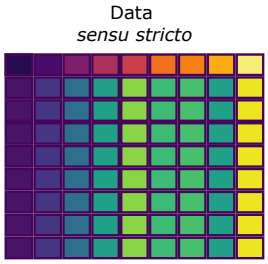
Thus, populating $\mathscr{B}_{Sensu\ \ stricto}$ may now be represented by Equation 5.8.

$$\mathscr{B}_{Sensu\ \ stricto} = \prod_{1,1}^{m,n} a_{m,n} \bullet \mathscr{T}_{Test} \bullet \mathscr{T}_{Trans} \tag{5.8}$$

Note that Equation 5.8 is inherently tolerant of elements $a_{m,n}$ that cannot be transformed because $\mathscr{T}_{test}$ does not return a value, since the matching $b_{m,p}$ element values are undefined. Using the alternative representation at matrix level is given by Equation 5.9, the completeness of transformation may be assessed by minimising the number of undefined elements in $\mathscr{B}_{Sensu\ \ stricto}$.

$$\mathscr{B}_{Sensu\ \ stricto} = \mathscr{A}_{Sensu\ \ lato} \bullet \mathscr{T}_{Test} \bullet \mathscr{T}_{Trans} \tag{5.9}$$

Data stakeholders may be underwhelmed by the production of data *s.s.* since this merely represents the holdings believed to be present from the start. Setting aside the need to verify the output for a moment, related data from the same template may now be shared and combined to facilitate analysis, so the production of data *s.s.* may eliminate the need for repeated manual manipulation of source data as part of an analytic task.



The data content are agnostic to format, but options include CSV, which may be imported into most analytic software. Thus, data *s.s.* may be incorporated into the stakeholder's existing analytic flow if the template is appropriately designed. However, the use of unverified data transformation is at best unwise, and while the mathematics has guided the overall reusable template concept, we must rely on data context to inform a verification process.

Example 5.3: Data sensu stricto.

We now have a mathematical description of linking the three states of data, which may be expressed in the simplified form of Equation 5.10 where the arrows represent transformations applied to matrix representations of data. Generally, the mathematical approach used here will apply to any data where such transformations may be defined.

$$\mathscr{D}_{Nascent} \implies \mathscr{A}_{Sensu\ \ lato} \implies \mathscr{B}_{Sensu\ \ stricto} \tag{5.10}$$

5.4 Template Concept

The essential characteristics of a reusable template built using the pseudocode of Figure 4.5 are represented in Figure 5.4. These have been creatively proposed as theoretical models using an approach that is justified under the Critical Realism umbrella as a novel theory. The figure shows multiple instances of same reusable template are linked by a dashed line to illustrate repeat use in different contexts. It is important to understand that this represents **multiple** applications of a **single** template with many inputs of related data *s.l* outputting multiple sets of data *sensu stricto*. The 'Reproducible Analysis' accepts data *s. s.* as its input for processing, however, it must be emphasised that the output requires contextual

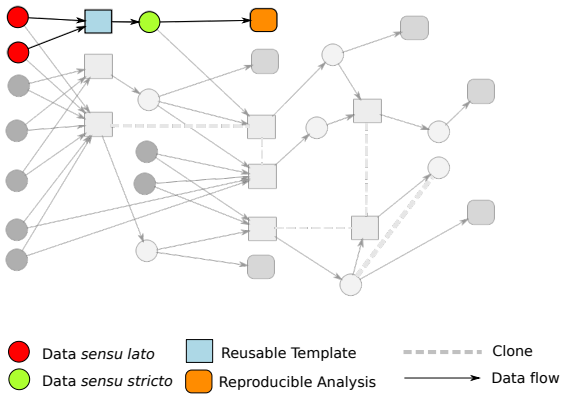


Figure 5.5: The basic template functional process is highlighted in this diagram. Multiple data *s.l.* are transformed by the template into data *s.s.* which in turn are the focus for analysis.

understanding to be interpreted as knowledge. For the purpose of generalising Figure 5.4, the 'Reproducible Analysis' may be thought of as reproducing the 'number crunching' element of an analytical task, prior to contextual interpretation. In the following paragraphs these characteristics are explored in more detail to validate the model by highlighting key parts of Figure 5.4 in turn.

Starting with the simple circumstance shown in Figure 5.5, two sets of data *s.l* are transformed into a set of data *s.s*, which in turn are the focus for analysis. A property that naturally flows from the matrix representation of data from Equation 5.2 and Equation 5.3 is the combination of multiple data instances by 'stacking' rows. We start to explore the useful implications of this property in Figure 5.6 where the same template is used to twice and combines multiple sources of data *s.l.*. Expressing this as a scenario inspired by current events (Brodie, 2020): Imagine that the seven data *s.l.*. are daily COVID-19 data-sets that have been combined as published as open data *s.s.*. If two more days of data *s.l.* are privately available, they may be combined with the open data *s.s.* and the reproducible analysis applied to investigate the effects of this additional data.

Staying with the same scenario as an explanatory tool, Figure 5.7 illustrates how the same data *s.s.* can be assembled by a researcher who only has access to groups of data *s.l.*, but the final combined data *s.s.* is identical in both Figures 5.6 and 5.7.

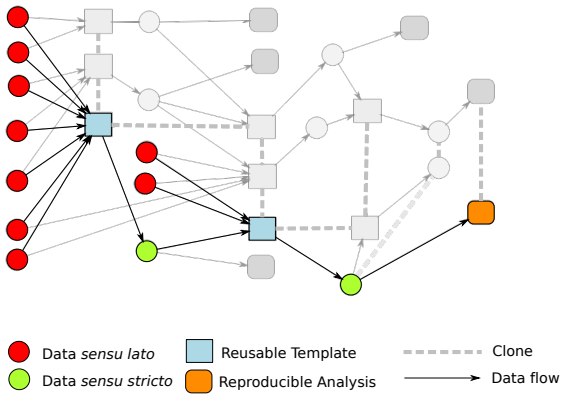


Figure 5.6: A property that naturally flows from the matrix representation of data *s.l.* and data *s.s.* is the combination of multiple datum rows.

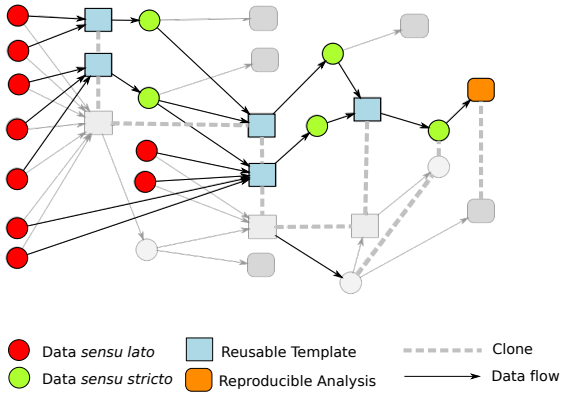


Figure 5.7: The order in which data are combined with templates does not matter, nor does the number of times that a template is used.

5.5 Template Implementation

The earlier inspirational work of Knuth (1984) has already been referenced. Knuth imagined that in the future researchers would be equally fluent in writing code and narrative into a 'web' document containing both elements. A web could be 'tangled' into executable code or 'woven' into a report. The key concept that has survived the test of time is the elegant mixing of code and narrative into a single document which may then be transformed by an external processor into multiple formats. More recently this same approach was described as a 'compendium' by Gentleman and Temple Lang (2007) who saw the combination of code, narrative and data as a method to allow reproducible confirmation of analytic results.

The reusable templates developed in this research are also text files that embody both narrative and executable code designed to implement the theory introduced in Section 5.4. The code and narrative are executed and woven together by an external program such as R Studio. However, the substantive difference is that the code elements are always written with reusability in mind and can, thus, be referred to as templates. For example, rather than loading named data files, they are 'discovered' by searching the data-raw directory as in Figure 7.4. This style of coding requires more effort in the early stages to accommodate unexpected conditions, but once developed, it allows for rapid iterations. Visualisations are managed in the same way, leaning heavily on the R package **ggplot** and its implementation of 'grammar of graphics' (Wilkinson, 2010) to produce well formatted visual outputs with minimal manual intervention. The coding style used to implement transformations implements the mathematical theory developed earlier in this chapter.

The compact representation of Equation 5.9 relates directly to the reproducible template of Figure 5.4 and provides justification for asserting that such templates are feasible constructs. The portion of Equation 5.9 given by Equation 5.11 eloquently captures the reusable template functionality in a form that guides the practical implementation using a programmatic language as conditional tests and transformations, with a definition of what happens when no matching test is found.

$$\mathscr{T}_{Test} \bullet \mathscr{T}_{Trans} \tag{5.11}$$

The property of row-wise combination of data *s.l.* allows us to treat multiple instances of

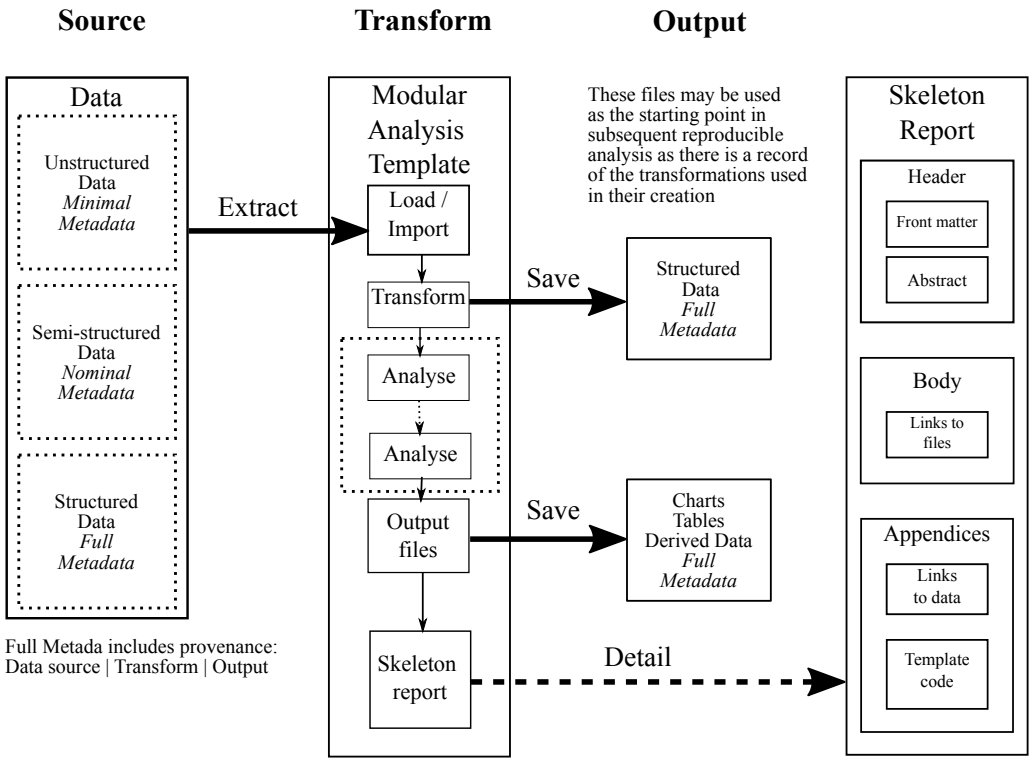
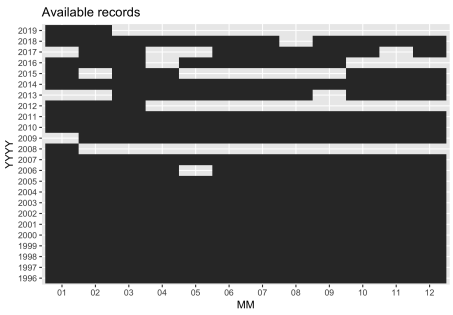


Figure 5.8: Modular analysis template concept.

$\mathscr{A}_{Sensu\ \ lato}$ as a single larger instance for the purpose of applying Equation 5.9. Since it is also possible to combine row-wise $\mathscr{B}_{Sensu\ \ stricto}$, all the implied template characteristics in Figure 5.4 may be met by appropriately combining $\mathscr{A}_{Sensu\ \ lato}$ and $\mathscr{B}_{Sensu\ \ stricto}$.

From equation 5.3, failure to successfully transform an element is assigned an undefined value which, although abstract is easily tested within $\mathscr{B}_{Sensu\ \ stricto}$. We once again return to the motivational example and in Example 5.4 consider the pragmatic application of Equation 5.9 with real data where a suitable transformation may not have been defined for each form of actual content. Since we are working with data too large to manually explore, it is clear that some form of summary must be included within practical templates to indicate successful transformation of elements. An appropriate descriptive analysis from Table 3.4 is included in Example 5.4 which again emphasises the need for understanding data context in the interpretative process. In the following chapter this theory is developed into a practical implementation using the R analytical language.

An initial descriptive analysis, guided by *a priori* knowledge of the data, should match inductively reasoned characteristics. In this case, the data were presented as Excel files containing many sheets, visible and hidden, that represent a complete record of the monthly observations over 40 years. When the output data *s.s.* are visualised as a tiled 'heatmap' of x-axis months and y axis years, there are gaps in the records which only go back to 1996, rather than 1980, which was a complete surprise to the stakeholders.



This serves to illustrate the importance of providing some basic visualisation of template output along with actual transformed data. All the data provided had been successfully incorporated as none were left marked as undefined. In this case the remedial action suggested is the search for more raw data source files but without having to manually search for specific content due to the number of files.
The question of duplicate data naturally arises in the conversation when searching for more nascent data sources to include since spreadsheets may be organised internally as multiple sheets into a 'Workbook'. Several such Workbooks may contain duplicate sheets included by users for analysis. As R implements a rich functionality to address this problem, the issue becomes one of selecting which fields are tested when considering duplicates. Closely related, is the issue of synonyms in otherwise identical records. Names of both species and locations change through time. However, defining preferred terminology and duplicates are context related choices. In this case, the Natural History Museum (2017) species dictionary was used to resolve synonyms to a current preferred name. As a separate task, a location dictionary was assembled to match older place names to those in current use. Finally, *date*, *place*, and *species*, were tested for defining duplicates.
Referring back to Figure 4.5, Applying transformations to attribute values before duplicate identification, allows these issues to be consistently resolved.

Example 5.4: An initial data verification though appropriate visualisation helps to confirm the effectiveness of the transformation process and data content.

The theoretical framework for reusable templates constructed has so far remained program language agnostic by using symbolic abstractions in the form of equations. This has effectively deferred the formal selection of a language until now. Chapter 2 noted the reasons for the popularity of the R analytic language for data intensive analysis, and as it had already been identified by the data stakeholders as a useful tool (See Appendix D), choosing the R Studio ecosystem of software offered an advantage in terms of stakeholder acceptance. This choice was validated in Section 6.7 by asking a stakeholder to remotely download relevant development template which was dependent upon a correctly installed R Studio ecosystem.

The motivational example serves to continually remind us that stakeholders are seeking to analyse data for specific goals, and are much more interested in repeatability of analysis with updated data than the underpinning theory that makes repeatability possible. A general purpose stakeholder orientated reusable template concept is presented in Figure 5.8 which illustrates all the functional components that are needed to support reusability. Stakeholders are not expected to interact with the transform elements at the code level, although the intention is to write in a coding style that supports end user modification. Rather, stakeholders will be given the opportunity to use the template in conjunction with their own data in a series of development micro-cycles, that implement feedback and suggestions. The novel template theory developed here describes the transformation of data through several states into data *s.s.* ready for analysis, and once this data state has been achieved, the process is well described. In Figure 5.4 data *s.s* is tidy, well-formed and may be regarded as the starting point for repeatable analysis.

5.6 A Practical Template Using R

The work by Cone (2018); Wickham and Grolemund (2016) has done much to promote the use of R, RStudio and R markdown as a literate programming environment supporting development micro-cycles by running short code chunks interactively and making amendments before committing to a full compilation. These techniques are well described in texts on R, but it should be emphasised that while there is much said about reproducibility, there is currently little said about reusability. However, we will demonstrate that the features of Figure 5.8 may be implemented in this environment, and that stakeholders successfully incorporated such implementations into their existing workflow.

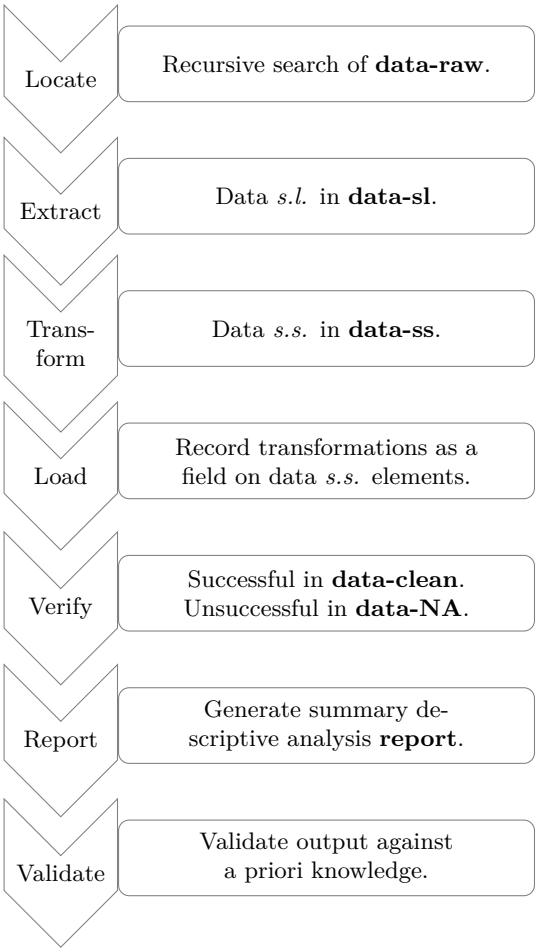


Figure 5.9: Core reusable template functional flow. Each core step is component in a practical reusable template based on the theory. Text marked in bold indicates physical directories where working files are placed.

While the details of every analysis will differ, Figure 5.9 illustrates the core functional flow that is common to all applications of the reusable template theory. The purpose of this diagram is to provide a high level guide to programmatic blocks of code that equate to the concepts presented in Figure 5.8. Physical directory names are provided to introduce a pragmatic standardisation of intermediate file location. In a perfect world where code works first time such conventions would have little utility, but in the real world where code may interact in unexpected ways with real data, the ability to examine each stage in the functional flow provides insights that can aid the code debugging process.

The choosing to split the functional flow into a single or multiple sub templates is application

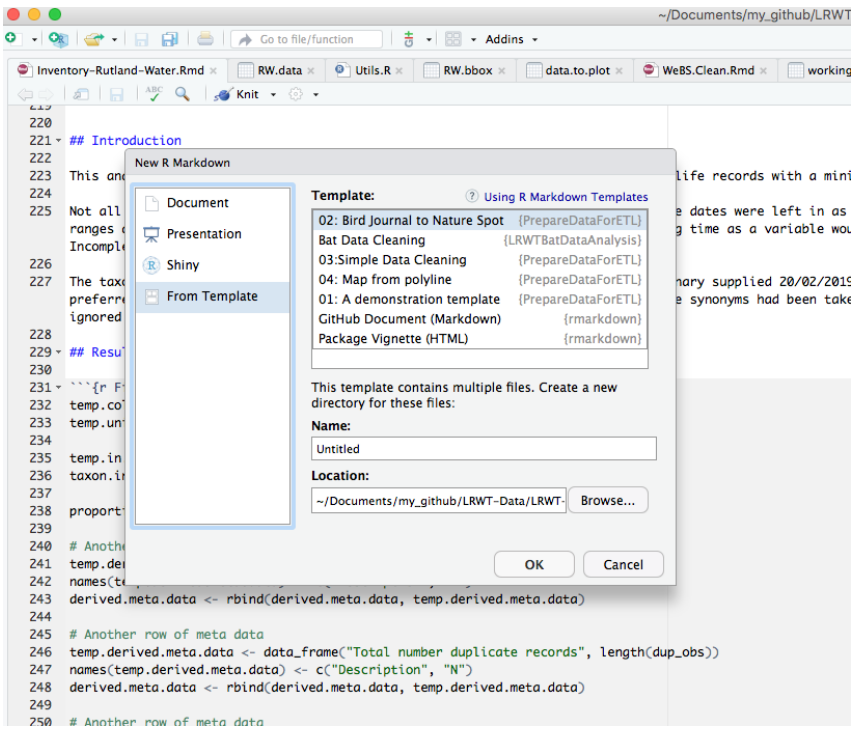


Figure 5.10: RStudio supports the inclusion of user selectable templates in packages as an installable extension to base R.

dependent dictated largely by time to run. The author's choice was to work with code elements that took a maximum of two minutes to run during development. This was highly dependent upon source data format and size, but at times during the course of this research up to 30 million rows of data were read for inclusion into templates. Typically, the initial load stage was slowest, so it proved convenient to produce data *s.l.* with one template and complete the process to data *s.s.* with a second.

The paths highlighted so far in Figure 5.4 have carefully avoided duplicate observations within data, but allowing the combining of multiple data implies that the reusable template will see data *sensu lato* that contains duplicates, thus requiring the definition of a functional behaviour to address this issue.

The issue of duplicate data *s.l.* is explored in Figure 5.12 where we see that combining data *sensu lato* should always result in identical data irrespective of the permutation data are transformed and combined through the reproducible template. If duplicate data are simply included then the length of data will depend on the path through data and the number of

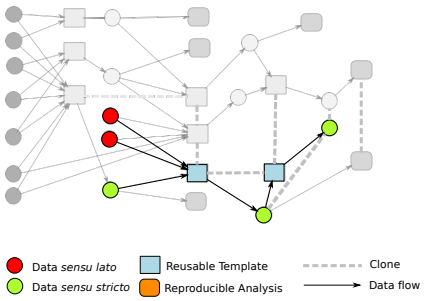
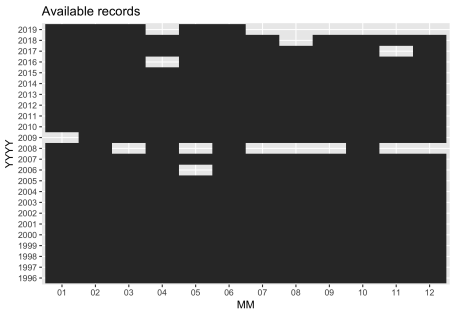


Figure 5.11: Combining data *sensu lato* and data *sensu stricto* for reproducible analysis.

The required template characteristic is shown in Figures 5.6 and 5.7 which illustrates how multiple existing data *s.l.* may be combined to provide a convenient reusable set of data *s.s.*. Combining this with two new sources of data *s.l.* found on an external drive provides a much better coverage, but still nothing earlier than 1996.



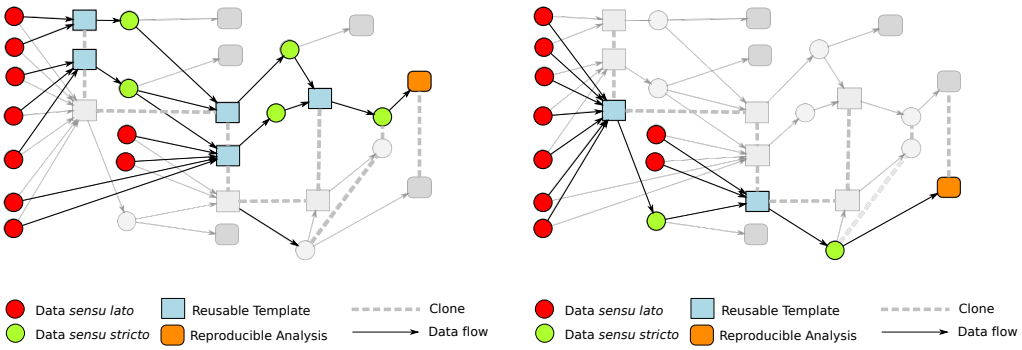
This example serves to show the importance of incrementally adding data *s.l.* sources without the need to keep rewriting template code.

Example 5.5: Combining multiple sources of data is an important characteristic of templates.

duplicates. With Big Data this may act as an unacceptable multiplier on data size.

A better approach is to introduce a metadata field that is used to describe characteristics such as source, transforms and duplicates relating to the data as a single entry. Thus, the multiple paths shown in Figure 5.12 result in identical data *sensu stricto*, but slightly different metadata, capturing the transformations applied. Refer to Example 5.6 page 83 see this in action. While this refinement is not strictly necessary for any single analysis, it keeps a permanent link to the source at the record level available for an as yet unspecified analysis.

RStudio supports the inclusion of user selectable templates in packages as an installable extension to base R, as illustrated in Figure 5.10, ensuring that dependencies and custom functions are also installed. Enabling this capability requires that all code passes the



(a) Five template passes.

(b) Two template passes.

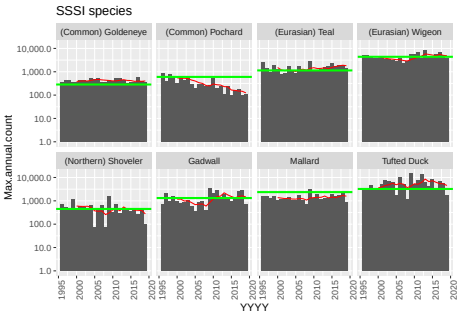
Figure 5.12: Different paths through the diagram produce identical data for reproducible analysis.

The available data are in many files and subdirectories that may be used in a fairly *ad hoc* manner. The stakeholders do not want to worry about duplicate observations within Excel files they just want clean data that can be analysed for management purposes.
Figure 5.12 shows how it really does not matter what order data are combined, and that final data *s.s.* depends only on the source data *s.l.* and is independent of the number of passes through the template.
This also impacts the user interface for the template, since stakeholders just want to get the job done, rather than spend time searching through the files to work out which ones need to be included using as 'select' type GUI interface. A 'data discovery' type interface that searches through a directory would be a preferable choice given the disparate nature of the data.

Example 5.6: Combining all data.

validation checks provided by R Core Team (2020). This point is made here in part to retrospectively justify the choice made for implementation, and in part to explain why no user interface was developed: it was already there. What was missing was a coherent theory of how to implement reusability, which was provided by this research. As a final technical point, templates may embed the code that call the required libraries directly and ensure that they are correctly installed, as such, they may be circulated as stand-alone entities and independently extended by programmatically skilled users. Data stakeholders focus on the outputs provided by Figure 5.8 and Example 5.7

For management purposes data are summarised on a rolling five-year rolling average, shown as a red line in the diagram. The green line is a value that indicates nationally significant numbers are present. The missing data remains a problem, but can be deferred without impacting the accuracy of current reports. These are used to judge the effectiveness of habitat management and changing external environmental factors and will have implications on future funding. Final validations of this report are made by the stakeholders comparing field notes with the summaries and external data sources. In this case they concluded that the report is an accurate summary of the source data, and that the template has saved many hours of manual data manipulation in a spreadsheet environment. Equally important, the familiar spreadsheet software may still be used to transcribe the monthly records and incorporated by the template to update the report on demand. Expert stakeholders may now focus on the interpretation of these observations and plan interventions that may be required.



Example 5.7: A stakeholder report illustrating the complexity of data visualisation required for some groups of species.

Chapter 6

Evaluation

The methodology in Section 3.7 noted the duality between user goals to improve the analytical process, and research objectives investigating stakeholders willingness to adopt new analytical tools. An indirect verification process was followed by working alongside stakeholders on current projects with a major analytical component. This approach enabled the evaluation from two viewpoints: verification of the template concept; and validation of the analytical method and results by stakeholders. As described in the methodology, stakeholders are motivated by the results, not by the academic research behind the methods used to achieve them. This is not to say that anything other than the highest accuracy is sought, but that, from their viewpoint, if timely delivery may definitely be achieved though manual data manipulation, that may be preferable to risking precious time on learning new tools that may not lead to success. Stakeholder validation is therefore based upon indirect evidence from the acceptance of analytic outputs provided by this research, rather than the creation and use of user templates.

Given the rapid accumulation of biodiversity data using electronic devices in the field, tasks where frequent report updating are required may be an attractive area for reproducible templates as manual techniques are too labour-intensive. However, the key contributions of this research are not the production of these reports, as the techniques used are well documented by the data science community, but rather, the challenges of assimilating nascent (or raw) data into the analytic process, which are often dismissed as data 'cleaning' issues due to imperfect data. This research has shown that an alternative viewpoint that considers the

process of data assimilation as one of changing data states leads to a versatile mathematically justified process supported by a sound underlying theory. This evaluation therefore begins by verifying the template concept with real data, before moving on to validation of the outputs from a stakeholder perspective. Referring back to Figure 5.8, the verification focusses on the Source, Extract, Transform blocks, and validation on the Output block.

6.1 Proof Of Concept

As proof of concept, templates were written to test the R Studio software ecosystem as a development environment. The design principles followed were those developed in Chapter 5 and also described in the draft paper of Appendix F.1. The initial data lepidoptera observations in Leicestershire were selected because the size and Excel format had become cumbersome to manage. With nearly 750,000 rows it was slow to load, and with the inclusion of new and historical records expected to exceed the Excel 1,048,576 row limit in the near future (Microsoft, 2020). Successfully demonstrating a template approach for reading and analysing the data was expected to provide credibility for deeper engagement with this research. Describing the data in terms of Shneiderman categories described in Chapter 2 it comprised of: 1-D; 2-D; Temporal; and Network attributes. Each of these had multiple issues with format, synonyms, content, although each observation was believed to be accurate. Conventionally one would start with 'cleaning' whereas the template approach instead seeks to transform the nascent data as provided into a self-consistent data *s.s.* with an intermediate data *s.l.*. The ease of reading the data allowed for rapid exploration and uncovered evidence of a previously unobserved phenotype of a common species of moth. The summary image in Figure 6.1 shows the day of year (DOY) by year for records of this species overlaid with k-means clusters for phenotypes. Stakeholders were far more interested in these results, rather than the method used to obtain them, which became the basis of a draft journal paper listed in Appendix F.3.

A second proof of concept test assembled a dataset of arachnid observations with similar internal issues into a 'book' style colloquially called an atlas within the biodiversity community. This was chosen to demonstrate the potential capability to frequently update such documents as they are usually produced in a labour-intensive process. The last updated atlas for arachnids was produced 2001, and the draft piqued interest from stakeholders. A

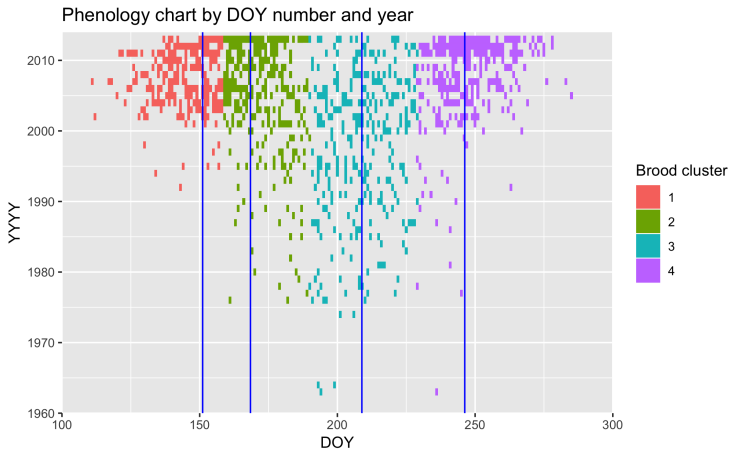


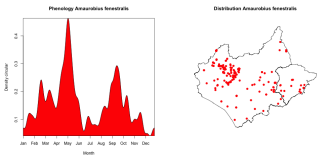
Figure 6.1: The proof of concept template was effective and uncovered evidence of a previously unobserved phenotype of a common moth species. The Brood clusters were identified statistically with the k-means algorithm and geospatial plots suggest that those labelled 1, 2, and 3 are able to exploit a wider range of habitats than cluster 3.

single page is shown in Figure 6.2. While this demonstrated the feasibility of technique, it also highlighted issues with the inclusion of non utf-8 symbols and unconventional use of characters in strings that can cause problems parsing data.

The lessons learned from these proof of concepts influenced the coding style adopted. Rather than silently berating the stakeholders for providing poor quality data, the starting point always assumed that such issues were present, even if they were not, so their sudden occurrence in updated data did not cause problems. In R, base functions include `make.names()` specifically to address such issues. It is the author's contention that reusability invites use with data of unknown provenance so code should robustly manage problems with data.

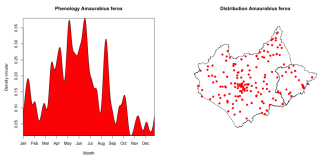
The conceptual visualisation of data proposed in the Methodology Chapter 3 Figure 3.7, proved to be an accurate representation, with the rectangular format of data worth special note, because it supports generalisation of analysis, as discussed in Section 7.1. To clarify this comment with an example, Lovelace and Cheshire (2017) describe how to create special 'SpatialObjects 'for manipulating geospatial data. This introduces a layer of complexity in that the need to convert back into a rectangular format quickly arises if any further analysis is required. An alternative, more versatile approach in the same situation, is to use 'Simple Features' to standardise the data into a rectangular format as described by Pebesma (2018). Network data may also be represented in the same way (Tyner et al., 2017). Rectangular

Amaurobius fenestralis



First seen: 1953 Last seen: 2014 Number of records:230

Amaurobius ferox



First seen: 1959 Last seen: 2013 Number of records:217

Figure 6.2: The proof of concept for a book format produced a 188 page book with custom front matter directly from records and auxiliary files.

data-frames in R may be easily saved as `.CSV` files allowing data *s.s.* to be passed for import into external software without the need for internal conversion.

The proof of concept was deemed successful, and the rectangular format of data adopted as the preferred strategy for saving data *s.l.* and data *s.s.* because of the inherent versatility of this representation. The biodiversity interpretation results were verified by expert stakeholders, so more focussed demonstrations of templates were produced.

6.2 Stakeholder Demonstration Templates

During interviews with stakeholders an inventory of species protected by nature reserves was discussed as an elusive goal. (See Apendix D.) This had been attempted by a manual approach using Excel spreadsheets but little progress had been made to problems with matching records to nature reserves and synonyms in species names. This last issue is common in biodiversity records as species may be split and merged over time. However, the UK Species Dictionary is managed by the Natural History Museum and they were happy to provide an electronic copy, which lists approximately 3 million UK species names in use, along with the current preferred name.

When reading the walk-through of the 'inventory of species' template steps described in Section 6.3, it is helpful to refer back to the generic pseudocode in Figure 4.5. The explanation follows the same terminology as in this figure, and is representative of the actual code used in the template implementation. Note that the 'Data Analysis' summarised in Section 6.4 is the stage used to assemble the final report and follows 'Tidy Data' (Wickham, 2014) principles. Wickham's approach includes refined tools for managing issues such as duplicate and missing data, and although a description of this well established techniques is outside the scope of this thesis, it is worth noting that the analysis is made reproducible because of the consistent presentation of data *s.s* from the source nascent data.

6.3 Pseudocode Description

Set up
- Load libraries and configure defaults. This is a housekeeping task specific to the analytical language used.

Data Discovery Search and loading of data;

- Species records;
- Geospatial data: nature reserves, county and district boundaries;
- Species Dictionary.

Data Representation The transformation of data into a form that can be used;

- Consistent representation of species records with standard dates, locations and preferred names.
- Flagging of records as inside or outside of boundary polygons. Note that records can be members of multiple polygons so this necessarily increases the number of fields in data *s.s.*

Data Validation Test plots and summaries of data *s.s.* used to check that it conformed to *a priori* expectations. Deviations were checked and the code modified as necessary.

Data Analysis Summarised in Section 6.4.

- Data *s.s.* contains all the derived fields required for a species inventory.
- Analysis consists of calculating summaries of the data and presenting in tabular form.

6.4 Summary Of Analysis And Related Issues

The LRWT provided a geospatial data set of their 45 reserve outlines and a complete set (629,305 rows) of their biodiversity record holdings. A two stage process was adopted to achieve a consistent dataset for analysis, and the results presented to the LRWT conservation committee. (See Appendix D.4.) The analysis revealed 6,718 unique species in the data, of which 5,327 had records which fell within reserve polygons, indicating the LRWT provides safe haven for 79% of species. After correcting for synonyms, 351,838 records were noted to

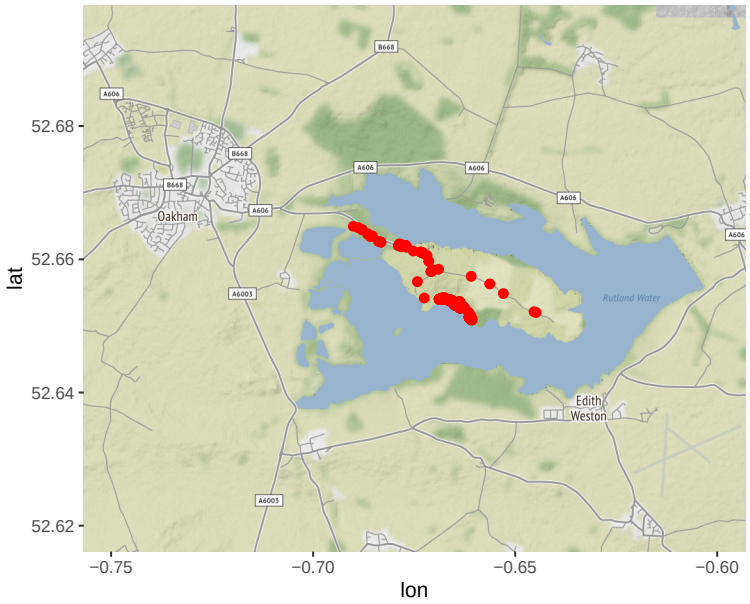


Figure 6.3: User controlled zoom to reuse the same code snippet in different geographic locations.

be duplicates, which was said to be in line with stakeholder expectations. The management of synonyms and duplicates are important issues for the biodiversity community, and while the results demonstrated that R could address them, the difficulty of learning to program was seen as a barrier to adoption.

Access to raw data prior to inclusion in central holdings was granted to explore the application of reusable templates in regular analytical tasks. The data relating to observations of bats provided a particular challenge. Sound recording from specialist equipment are annotated by experts and these are transformed into summaries. Multiple type of equipment are used and the type of output file vary. Data are stored by type, rather than by survey, so a template must cope with varying file types, formats and annotations. This template implemented three innovations: A custom package created using `utils.template.package.creator` ; private data separated from package development (See Figure 7.4); user customisable mapping (See Figure 6.3.)

The speed with which these demonstrations were produced elicited requests for custom maps and the transformed data s.s. in `.CSV` format from (more or less) standard survey spreadsheets. These spreadsheets effectively capture ecological observations in a 'wide' format, but analysis requires the 'long' format of data *s.s.*, a transformation task currently achieved by

manual manipulation of spreadsheets. Using the generated `.CSV` files save many hours of work and was easily incorporated into the established workflow.

6.5 Verification

The core reusable template functional flow is shown in Figure 5.9 which is a practical implementation of Equation 5.9 a consequence of the linear nature of the process is that the output in data-clean comprises data that were located in data-raw and not rejected in data-NA. Tests to confirm the size of located contents are context related, but generally, the consequence of not locating data are trivially obvious, so ensuring data-NA remains empty is the primary goal when designing a new template. When the template is running, the descriptive analysis report is the primary verification tool. This concept was introduced in the motivational example, and here, Figure 6.4 demonstrates the principle in a more complex situation. These are labelled to be approximately 270, 000 biodiversity records pertaining to Leicestershire, but a basic geospatial representation shows the presence of records from all over the UK necessitating 'clipping' to the area of interest. A publicly available outline of Leicestershire (National Biodiversity Network, 2018) was used to generate a bounding box that could be used to crop the data to the subset that could be plotted on Figure 6.4. A convention with such data is to summarise over 2 kilometre squares (tetrads), which necessitates the re-projection of all the data into a common coordinate format prior to plotting. Here, a logarithmic scale has been chosen to represent the record count per tetrad and zero counts, which are undefined on logarithmic scales, are shown as transparent.

The stakeholders were unaware of the 'out of area' records which could have caused unpredictable issues, but otherwise the check plot confirmed successful transformation of data. Clearly, there are issues relating to data distribution that will need to be addressed in any analysis, but the template has done its work and will be able to transform additional data as it becomes available.

Not all transformations can be verified with a single simple plot and it may be that they uncover issues that require further investigation. An example is from a feasibility study: Figure 6.5 where the minimum size of habitat classifications are <100 m^2 which seems rather small given the context. Manual checking is not feasible given that the base data comprises approximately 30 million rows, but a sample check suggests that the linear dimensions of

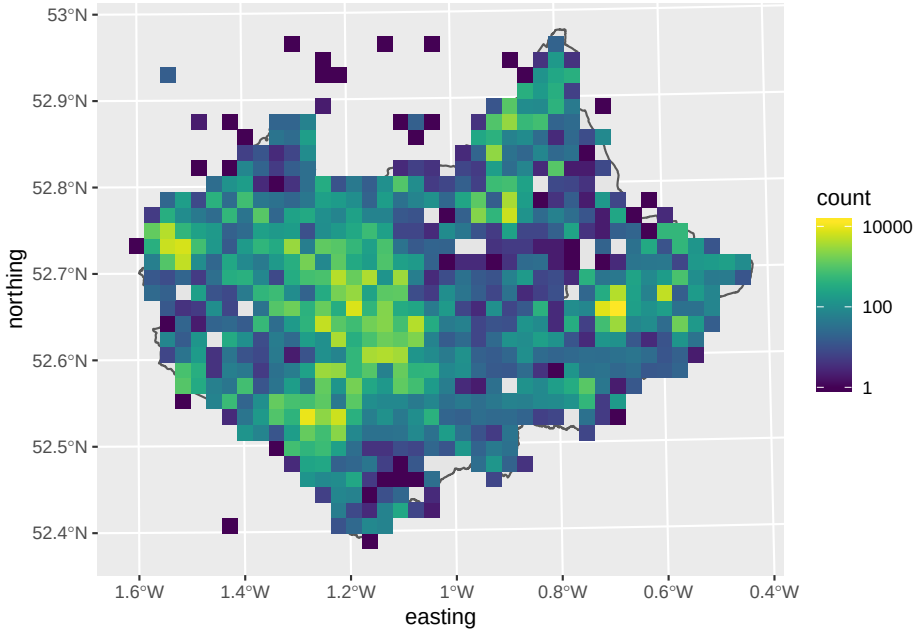


Figure 6.4: Complex check plot merging data from multiple sources.

polygon edges and the enclosed area are not correctly calculated, resulting in the small areas observed. A brief investigation suggested the root cause was an incorrectly applied global rescaling of source data in the QGIS mapping software used to collate data from multiple sources. With the caveat that further verification was required an approximate correction was applied to enable the feasibility study to proceed.

6.6 Verification Summary

The selection of suitable descriptive summaries to verify transformation provides a qualitative indication that an expectation based on *a priori* knowledge has been met. Where this is not the case, as in Figure 6.4 with 'out of area' data, a suitable intervention will need to be applied. Typically, the output will only be shared once reasonable expectations have been met, so there is an element of survival bias in this approach. Although in some cases it may be appropriate to apply a scaling factor to allow the project to continue, the viewpoint remains one of data transformation, rather than cleaning of faulty data. Generally, the mathematics of templates were found to work robustly with biodiversity data and proved a valuable guide when creating templates with the characteristics of Figure 5.4. The

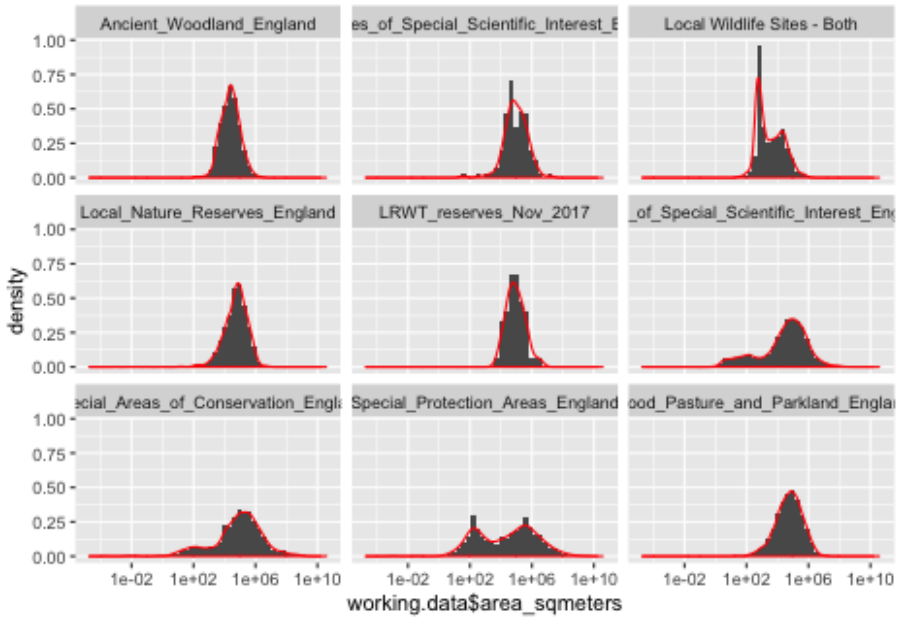


Figure 6.5: Check plots may reveal complex issues with data even though the transformation is correct. In this plot of the area of geospatial polygons describing habitats, the small size, <100 m^2, needs investigation.

Figure 5.9 and was an accurate description of the process flow.

The modular analysis template concept of Figure 5.8 was implemented using the R Studio, Markdown and KnitR ecosystem of software. The development of template theory enabled the reproducibility of this approach to be made reusable. This replaced a manual transformation of data usually accepted as necessary by stakeholders, as described in Section 4.1.

6.7 Validation

Early in this project it was confirmed that stakeholders were willing to adopt R Studio into their workflow and that they were to download and run CRAN-R compliant package and demonstration template. Instructions published on the software development hub `github.com` [1] were successfully followed remotely by a motivated LRWT stakeholder with the approval of organisational IT support services. This was regarded as validation of the software platform chosen for implementation of templates, and confirmation that stake-

[1] `https://github.com/enpjp/PrepareDataForETL/wiki`

holders were prepared to expend intellectual effort to advance their analytical capabilities. However, discussion about the theoretical framework for creating templates was deferred to future stakeholder workshops beyond the time-frame of this research. The changes in working practice brought about by the COVID-19 pandemic ended regular physical meetings abruptly, but the expectation generated from ongoing online meetings, are that long term engagement with the stakeholders will continue, but with greater emphasis on analysis as the barriers to accessing data have been reduced by this work.

Outputs were regarded by expert stakeholders as valid and informative. It is understood that some have been circulated to help inform environmental policy.

Observations made during custom surveys by professional ecologists are usually transferred to a 'standard' spreadsheet which is then customised to the meet specific requirements. Analysis requires manually transforming the data to create summary visualisations for incorporation into a client report.

A reusable template generated data *s.s.* were incorporated into stakeholder spreadsheet analysis saving many hours of work. Previously, the transformation of raw data for analysis was seen as an area for manual work with no opportunity for improvement, so this was seen as an unexpected benefit. It is possible that further benefits to stakeholders may arise if an analytic template were used to generate the template report too. However, the ecologists are highly skilled and once the roadblock of data transformation was removed, report development was rapid.

During the immediate time frame of the research project there was no independent adoption of the template into LRWT outputs. However, secondment by the researcher into the organisation was agreed, with the expectation that engagement would continue beyond this research project to assist with data analysis tasks. It was also clear that several members of staff were keen for workplace tutorials about R and R Studio to build skills in this area.

Liaison with Leicester University NERC project: NE/S009310/1 highlighted the problems with preparing raw data for inclusion in the analytic process. This project developed a tool for use with QGIS mapping software to overlay habitat information. The import functionality within QGIS was able to accept data *s.s.* from several reusable templates confirming that the block in the process was initial transformation, rather than the data availability. QGIS was used to generate exploratory views that could also have been generated through the

templates, and although this process was more time-consuming, it did not require programming skills to manipulate data through the WYSIWYG interface. When an analytical task is essentially a one off, it seems likely that stakeholders will tend to use familiar solutions under their control. However, there are projects currently being developed where regular summary documents are required and stakeholders have expressed a desire to use templates to create reports based on those used to dynamically build Figures 6.4 and 6.5.

6.8 Summary

The opening paragraphs of this chapter posed four questions for evaluating the output of this research. Extracts from working templates using stakeholder data confirm the functionality and robustness of this approach. The inclusion of a suitable descriptive analysis guides both the design of templates and the interpretation of data. The answer to the first two questions is: yes templates work; and yes, they cope with the unexpected quirks in real data. Answering the questions about meeting stakeholder requirements and long term usefulness is more nuanced.

Stakeholders had previously identified the R Studio platform as a potentially meeting their requirements but found the incorporation of raw data to be a manually intensive barrier. The applied template theory removed this barrier, but the skills to use the R analytical language were lacking, making continued use of QGIS and spreadsheets an attractive hybrid solution.

Procedures are in place for the LRWT organisation to accept volunteer help in many types of role. Thus, projects such as BOM(Biodiversity Opportunity Mapping) Lite, were happy to receive help in the production of high quality summaries for ecologist to incorporate into reports. It seems likely that the author will have a long term relationship with the LRWT helping to apply the outputs of this research.

Chapter 7

Discussion

Referring back to the research question: Can an empirical theory of the knowledge extraction process be developed that guides the creation of tools that gather, transform and analyse *nascent* data? The answer is yes, and the main barrier for uptake is the necessary level of programming skills. Using existing templates requires a much lower level of expertise so reuse is more likely when programming skills are limited to a few individuals. The secondary question: 'Will data stakeholders use these tools?', is less clear at this stage, and dependent upon the stakeholder group developing better R based programmatic skills, which they are keen to do. However, stakeholders were quick to import data *s.s.* `CSV` outputs into their existing workflow using familiar software tools, indicating that they perceived added value in the transformation.

A curious observation that arises from this research is how introducing an abstract concept 'state' imbues the data with properties that facilitate analysis without changing the information contained therein. The inductive process that suggested such a concept was born in Figure 5.4 by framing the problem around reproducibility and reusability, thus effectively expressing the desired behaviours and then exploring how this could be achieved. A clue to the explanation is provided by Figure 3.3 and the mapping of the template theory onto the data definition, where we might regard 'state' as ways of organising datum triples, but without changing them. Along with the mathematical expression of Equation 5.11 that describes the properties of reusable templates, the idea that the same information is presented in several states naturally arises in the mathematics as shown in Equation 5.10 where the

arrows represent transformations applied to matrix representations of data.

Thus, the differences between these three representations of the same data are in the *organisation* rather than content. The progression from disorganised to an organised state may be equivalent to entropy, and here we find agreement with other research. Entropy has been applied as a data cleaning tool by Yakout et al. (2011) and Chu et al. (2015) in broadly similar approaches that seek to minimise user interventions. This contrasts with the viewpoint of this work that user context is always required for correct interpretation of raw data. However, given that entropy may be a measure of data state it is a promising approach that might be incorporated into the template process developed here.

The link between data entropy and the reusable templates of this research are unproven, but the expectation would be that entropy decreases as the data are progressively more organised by expending 'work'. In this case the work is expended by the template. A unique idea introduced here has not been to quantify this change of state, although this might be academically interesting, but has been instead to make use of the properties that emerge as a result of the reorganisation. It is a small extra inductive step to suggest that there may be many combinational states that are true for $\mathscr{D}_{Nascent} \implies \mathscr{A}_{Sensu\ \ lato}$ and have similar values of entropy, but only those that meet Equation 5.2 have the required properties.

Justification for creating the model as a goal and then building a matching theory are consistent with the critical realism philosophy as explained by Williams and Wynn (2018). It was suggested that progress on many problems in physics and engineering could be made by constructing novel models and theories which are then tested against the observed world rather than trying to incrementally improved the *status quo* of accepted wisdom.

7.1 Generalisation Of This Research

Throughout this work the application has been relegated to a secondary role, as has the implementation, which has been explained using mathematics and diagrams, rather than specific computer code. The generalisation started with the choice of data definition used in Chapter 2 and presented in the frequently referenced Figure 2.2. It is argued here that the template approach will work for *any data where this definition is valid, and useful where the data cannot be conveniently handled by other methods.* The caveat is needed as effort is

required to create a reusable template, so it should not be seen as a universal panacea.

The theory upon which the template has been built is agnostic to the real world domain, as is shown in Figure 3.3, and any real word observations that can be represented as datum triples are equivalent to $\mathscr{D}_{Nascent}$ of Equation 5.1.

Generally, the mathematical approach used here will apply to any data where a matrix representation is applicable. From an R perspective, this equates to anything that maps to the data-frame entity which is the building block for the Tidy R approach pioneered by Wickham and Grolemund (2016). This mapping has been so successful that new tools are emerging to exploit the simplicity of the data-frame. For example, the Simple Features representation of spatial data developed by Pebesma (2018) encodes geospatial data (points, lines and polygons) in a way that simplifies its combination with any other data-frame representation. Networks may also be presented as a data-frame, which is the base entity used by `ggplot2`, the *de facto* standard tool for visualisation (Tyner et al., 2017). There is no evidence that any of the data types mapped on to visualisation charts in Tables 2.1—2.3 are not covered by the data definition of Figure 2.2, so all are compatible with the data-frame representation.

However, the data are not knowledge without an understanding of the context, which is why the research data, generically represented in Figure 3.7, needed to have a clear connection to the real world, and as justified in Chapter 4, the biodiversity community were top of the list. There are no reasons to suppose that the issues encountered with biodiversity data are in any way unique to the domain. In addition, examples cited by Bowker et al. (2013) in ‘ “Raw data” is an oxymoron (Infrastructures)’ cover economics, astronomy, trade databases, social science, internet records and biodiversity. Only time will tell if the techniques developed here will be useful in other domains, but incompatible data structure is not likely to be a reason.

7.2 Data Sharing And Provenance

The literature review in Section 2.9 noted several areas relating to data provenance that were not chosen for research at this time.

- How can transformations be recorded and verified?

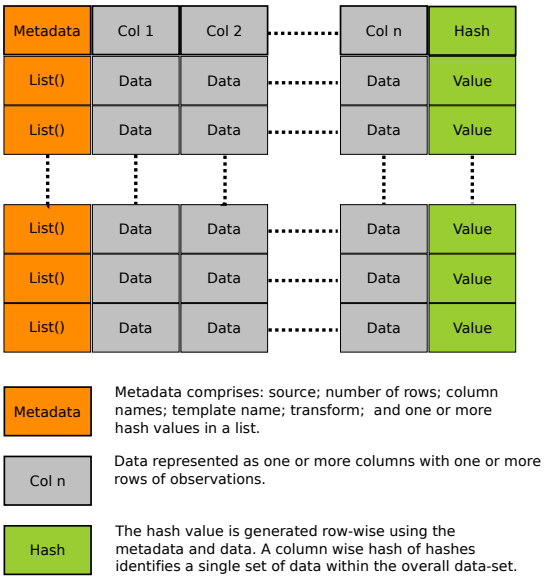


Figure 7.1: Principles for digitally signing data are well established.

- How can data provenance be assured?

References citing the importance of reproducibility and data sharing given in Section 2.6 suggesting techniques that support data provenance are likely to be of increasing interest to the research community. The natural flow of $\mathscr{A}_{Sensu\ \ lato} \implies \mathscr{B}_{Sensu\ \ stricto}$ provides an opportunity to record details of source and transformation a signed metadata field in $\mathscr{B}_{Sensu\ \ stricto}$ as shown in Figure 7.1.

This was implemented in an unsigned form in the templates created for this research as recording data source provided helpful information, especially during writing and debugging templates. The principles of digital signing are well established, however, there are a number of potential schemes possible and their use would impact the way in which templates are written and data shared, so further research is required to understand if this added level of complication would be effective in practise.

If data sharing to support reproducibility in research becomes more popular some system of signing to prove provenance will be required, especially if the data are too big for manual inspection. However, the best method by which this may be achieved is still an open question.

Additional possibilities for working with shared and private data are created by properties

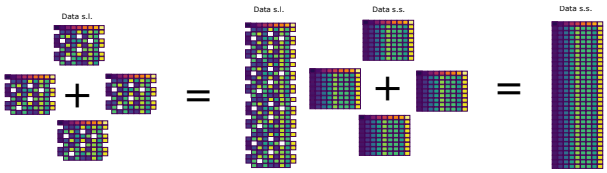


Figure 7.2: Shared data may be combined using the principles of data states. Data *s.l.* may be combined with data *s.s* using a template.

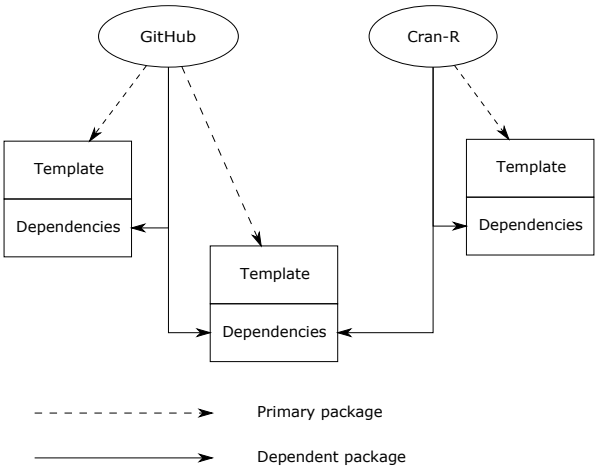


Figure 7.3: Templates may be shared though the package system within R. GitHub may be used to share 'development' packages that have not been through Cran-R verification. Missing dependencies always return a warning message, and may be manually or automatically downloaded according to template configuration and source.

of $\mathscr{A}_{Sensu\ \ lato}$ and $\mathscr{B}_{Sensu\ \ stricto}$ which are illustrated in Figure 7.2. Remembering that a template will transform $\mathscr{A}_{Sensu\ \ lato} \implies \mathscr{B}_{Sensu\ \ stricto}$ there is great flexibility in working with data from multiple sources. Thus, a 'private' and 'public' data may be combined for analysis without exposing the 'private' elements. If local data contains sensitive attributes there is a *prima facie* case for exploiting this principle.

7.3 Sharing Templates

The principle of sharing templates using GitHub and development packages were demonstrated early in the research process when a motivated stakeholder successfully installed a prototype template remotely using the online instructions provided at:

`https://github.com/enpjp/PrepareDataForETL`

This repository passes the technical requirements for publication on the formal Cran-R repository as demonstrated by the Travis CI (Continuous Integration) pass logo which is automatically created after each update. GitHub and CranR repositories may be mixed as long as dependencies on other packages are carefully managed. (See Figure 7.3.) While code writing to this high standard is desirable, it is not clear that it is appropriate to publish templates that are specific to particular data orientated problems as full packages. In part, this is because care is needed not to accidentally incorporate private information into the repository, and partly because packages are intended be used for extending generic functionality in R, so may be rejected by Cran-R for not meeting this criterion.

As experience in writing templates grew so did a pragmatic approach: generic functions were placed inside a development package and called as required from a template. This ensured complete separation of code from data, allowing the package to be publicly shared via GitHub. Creating packages is a tedious and unforgiving process when manually undertaken, however, the new tools developed by Wickham and Bryan (2015) allow the process to be incorporated into an R script. This in turn allowed the creation of a 'package to create template packages' specifically for this research:
`https://github.com/enpjp/utils.template.package.creator`
A particular feature is exploitation of the package 'inst' directory to create a place for template development outside the package build, and within its own private repository avoiding any risk of test data 'leaking' into the package history during development. A search using the term 'remove sensitive data from github' returns over 5 million hits suggesting that this is a common issue while developing software, so greater awareness is needed about the features available to manage this problem.

Figure 7.4 shows how the package 'inst' directory and subfolders are used for non-programmatic files and data. Two files, *.rbuildignore* and *.gitignore* control compilation into the package and inclusion into the git repository respectively. Correctly phrased stanzas are inserted by `utils.template.package.creator`. There is no doubt that this is a technically challenging implementation to achieve, however, the tools provided by Wickham and Bryan (2015) make this structure reproducible with only two commands issued through `utils.template.package.creator`. The code and directory structure created pass the R verification tests allowing for regular checking during development ensuring the quality of coding, leaving the author to focus on validation.

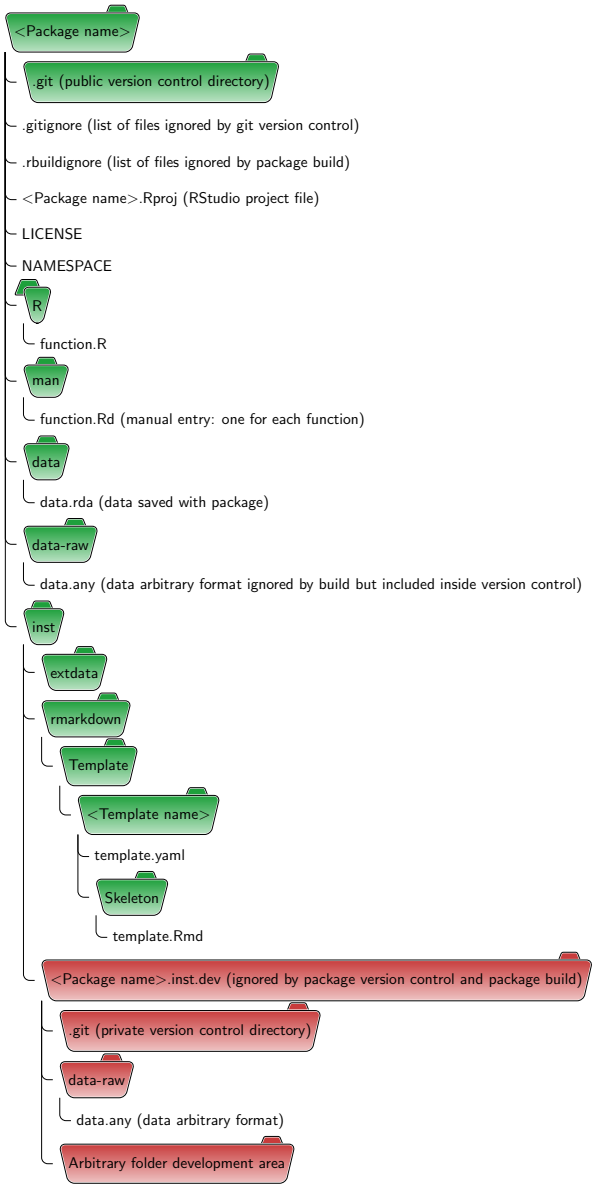


Figure 7.4: The template directory structure is specific to R packages and shows how the public package (green) and private development files (red) are separated and the naming conventions used. Correctly configured stanzas are inserted into *.rbuildignore* and *.gitignore* by `utils.template.package.creator`.

7.4 Analysis Of Data

As the template approach now unlocks $\mathscr{D}_{Nascent}$ for analysis, where to next? Authors such as Chen and Zhang (2014), make a good case for Data Intensive Scientific Discovery (DISD) and infrastructure for data management, but much less is said about the statistical tools required. For good reason, mechanistic analysis, as shown in Figure 3.4 is often cited as the 'gold standard' of scientific analysis, especially in clinical trials. But it is not always possible to set up controlled trials and then to collect the data. Causal analysis, as pioneered by Pearl and Mackenzie (2018), offers a formal route to separating cause and effect from observational data by using causal models to capture dependency between variables. While not a universal panacea, the causal analysis may be appropriate in many situations where data have been collected in advance of setting an analytic goal, and also indicate situations where the data are inadequate to support the desired analysis. This neatly moves the focus of this discussion onto analytic challenges which are the primary motivation of data stakeholders.

The stakeholders who provided data for this research were experts in biodiversity with considerable experience in using spreadsheet and geospatial tools for analysis. However, the variety and volume of data was a choke point as it had to be manually transformed for use in analysis. So deeply normalised was the need for manual intervention, accessing the real raw data was often problematic as stakeholders always wanted to tidy it first. This observation supports the choice of title by Bowker et al. (2013) '"Raw data" is an oxymoron', and why effort was needed to access 'Raw data'.

Data *s.s.* when presented to stakeholders as a CSV file was compatible with Excel Spreadsheets and QGIS, so enabled template output to be rapidly incorporated into the established analytic and report writing process. During the course of this research, no stakeholder utilised templates independently, instead focus shifted to more ambitious analysis using data now made accessible using a template to produce data *s.s.* for import into other software tools. Coupled with a desire to provide evidence based biodiversity reports with sound statistical justification, causal analysis is a natural next step.

Chapter 8

Conclusions

This research has sought to answer two questions that were identified through the literature review of the Big Data domain:

> Is it possible to create tools that gather, transform and analyse *nascent* data? A secondary pragmatic question followed naturally from the first: Will data stakeholders use these tools?

These questions were explored using biodiversity data for two reasons: large amounts of raw data were available which presented many problems when analysis was attempted using existing approaches; and the data stakeholders were willing to share the data in return for analytical help. It is argued that the issues in biodiversity data are typical of many other sources where collation is antecedent to analytical definition.

The answer to the first question is yes: this research has demonstrated that reusable templates are an effective tool for incorporating high variety and volume data into the analytic process. This has been achieved by introducing a novel concept of 'state', justified by an underpinning theory, to guide transformation. The answer to the second question is less clear at this stage, and dependent upon the stakeholder group developing better programming skills. However, stakeholders were quick to incorporate template output into the existing workflow. It was notable that the need for manual intervention on raw data has become so normalised, that terminology such as 'Dirty data' biases thinking to see faulty data needing 'cleaning' as the obvious next step, rather than one of improving the analytical process, as

explored in this research. That is not to say that the excellent work that has been done in 'data cleaning' is not worthwhile, but instead that there are ways to work with raw data that do not start with a presumption of imperfection. This thinking underpins the theory developed here to propose states of data and leverage mathematical properties to guide the transformation into data *s.s.* with well-defined combinational properties that simplify the manipulation and analytical process. While the outputs have proved extremely useful to stakeholders, the technical skill required to write templates is a barrier to their use. Much less skill is required to run templates, and while stakeholder use was demonstrated under test conditions, they have not yet been used independently. Long term plans to work with stakeholders may encourage uptake on specific projects, but at the time of writing, this remains an ambition.

Analysis of data 'unlocked' by templates has been achieved. Two papers, as yet unpublished, explore two biodiversity sets, and an internal report seeks to uncover 'hidden' opportunities for evidence based conservation interventions. Literate programming techniques are complementary to the template approach, and as these tools are now so well-supported in the R Studio and Markdown ecosystem these are the natural way forward for implementation.

8.1 Novel Contributions

The novel contributions are within the theory of templates developed in Chapter 5 which also introduced the concept of reusability as an extension of reproducibility. This theory also sets the context in which reusability complements the existing literate programming techniques that underpin reproducibility: essentially defining reusability as a subset of reproducibility. The task rather than data orientated approach implied by reusability, requires the introduction of new terminology, reproduced in Figure 8.1, to clearly convey the state of data in relation to the task. The mathematical framework describing the essential properties that are required for a template to be reusable is illustrated in Figure 5.4, and the framework and derived properties lead in turn to a systematic method for the functional implementation of reusability illustrated in Figure 5.9.

The Chapter 6 described an evaluation of the theory through empirical demonstration using a number of templates, each of which using the mathematical theory as a guide. Their creation was facilitated by developing a custom utility `utils.template.package.creator`

***nominal* data fields** This term is used to describe the **presumed** data fields and structure.

data *nascent* This is the **actual** initial state of data. Variances from the presumed nominal state are often described with pejorative terms such as 'messy' and 'untidy'.

data *sensu lato* Once transformed into a readable rectangular state, this raw data is termed as data *s.l.* to emphasise that data may need 'cleaning' or other transformation before use. Multiple instances of this state may be combined row wise to form a larger data *s.l.* set.

data *sensu stricto* Once data are transformed into a fully defined state ready for analysis it is termed data *s.s.* . Multiple instances of this state may be combined row wise to form a larger data *s.s.* set. However, if any data *s.l.* are included in such a combination, the result are data *s.l.*.

Figure 8.1: A novel terminology of data states. Note that changing the state does not create or destroy information.

that creates a skeleton reusable template framework to cran-R standards, automating an otherwise technically unforgiving 20 + step process.

The implementation used R Studio as an IDE and followed R package conventions to manage documentation. All code was verified locally using `devtools` to meet cran-R requirements for formal publication as a user contributed extension package. Development version of templates were published on public and private Github repositories along with supporting documentation ensuring separation of code and data, as described in Section 7.3. Repositories used Travis CI (Continuous Integration) to ensure that the Gihub version meets cran-R requirements after the application of updates. The routine observance of following these formal standards allowed the successful sharing of template packages via the internet using established GitHub and R development package support services.

References

Abbott, B. P., Abbott, R., et al. Observation of Gravitational Waves from a Binary Black Hole Merger. *Physical Review Letters*, 116(6):061102 (2016). ISSN 0031-9007. doi:10.1103/PhysRevLett.116.061102.

Akiwatkar, R. The Most Popular Languages for Data Science - DZone Big Data (2017). **URL:** *https://dzone.com/articles/which-are-the-popular-languages-for-data-science*

Ali, S. M., Gupta, N., et al. Big data visualization: Tools and challenges. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 656–660. IEEE (2016). ISBN 978-1-5090-5256-1. doi:10.1109/IC3I.2016.7918044.

Allaire, J. J., Xie, Y., et al. rticles: Article Formats for R Markdown (2019).

Almaliki, M., Ncube, C., et al. Adaptive software-based Feedback Acquisition: A Persona-based design. In *Proceedings - International Conference on Research Challenges in Information Science*, pages 100–111. IEEE, Athens, Greece. (2015). ISBN 978-1-4673-6630-4. ISSN 21511357. doi:10.1109/RCIS.2015.7128868.

Andor, C., Joó, A., et al. Galois-lattices: A possible representation of knowledge structures. *Evaluation in Education*, 9(2):207–215 (1985). ISSN 0191765X. doi:10.1016/0191-765X(85)90015-1.

Ayto, J. and Crofton, I. Wrong sort of snow. In Ayto, J. and Crofton, I., editors, *Brewer's Dictionary of Modern Phrase & Fable.* Chambers Harrap Publishers, 2 edition (2009). ISBN 9780199916108.

Azzone, G. Big data and public policies: Opportunities and challenges. *Statistics and Probability Letters*, 136:116–120 (2018). ISSN 01677152. doi:10.1016/j.spl.2018.02.022.

BEIS. Companies House (2018).
URL: *http://download.companieshouse.gov.uk/en_output.html*

Beyer, M. A. and Laney, D. The Importance of 'Big Data': A Definition. Technical Report June, Gartner, Stamford, CT. (2012). doi:G00235055.

Bhaskar, P. R. *A Realist Theory of Science*. Verso, London, UK (2008). ISBN 9781844672042.

Big Data Public Working Group. NIST Big Data Interoperability Framework: Volume 1, Definitions NIST Big Data Public Working Group Definitions and Taxonomies Subgroup. Technical report, NIST (2018). doi:10.6028/NIST.SP.1500-1r1.

Billestrup, J., Stage, J., et al. Persona usage in software development : Advantages and obstacles. In *Conference on Advances in Computer-Human Interactions*, pages 359–364. Barcelona, Spain (2014). ISBN 9781612083254.

Bivand, R. and Krivoruchko, K. Big data sampling and spatial analysis: "which of the two ladles, of fig-wood or gold, is appropriate to the soup and the pot?". *Statistics and Probability Letters*, 136:87–91 (2018). ISSN 01677152. doi:10.1016/j.spl.2018.02.012.

Boccaletti, S., Bianconi, G., et al. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122 (2014). ISSN 03701573. doi:10.1016/j.physrep.2014.07.001.

Bowker, G. C., Brine, K. R., et al. *'Raw data' is an oxymoron(Infrastructures)*. MIT Press, London, UK (2013). ISBN 978-0262518284.

Boyd, D. and Crawford, K. CRITICAL QUESTIONS FOR BIG DATA. *Information, Communication & Society*, 15(5):662–679 (2014). ISSN 1369-118X. doi:10.1080/1369118X.2012.678878.

Brodie, M. PHE statement on delayed reporting of COVID-19 cases - GOV.UK. Technical report (2020).

Callaghan, S., Donegan, S., et al. Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *International Journal of Digital Curation*, 7(1):107–113 (2012). ISSN 1746-8256. doi:10.2218/ijdc.v7i1.218.

Candes, E. About the Replicability of Scientific Research in the Big Data Era: What Statistics Can Offer? In *Le¸ cons Jacques-Louis Lions* (2017).

Cavanillas, J. M., Curry, E., et al. *The Big Data Value Opportunity*, pages 3–11. Springer International Publishing, Cham (2016). ISBN 978-3-319-21569-3. doi:10.1007/978-3-319-21569-3_1.

Ceri, S. On the role of statistics in the era of big data: A computer science perspective. *Statistics and Probability Letters*, 136:68–72 (2018). ISSN 01677152. doi:10.1016/j.spl.2018.02.019.

Chang, A. The Facebook and Cambridge Analytica scandal, explained with a simple diagram (2018).
URL: *https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram*

Chen, C. P. and Zhang, C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275:314–347 (2014). ISSN 00200255. doi:10.1016/j.ins.2014.01.015.

Chen, H., Chiang, R. H., et al. BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT. *MIS Quarterly*, 36(4):1165–1188 (2013). ISSN 07308078. doi:10.1145/2463676.2463712.

Chu, X., Morcos, J., et al. KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*, pages 1247–1261. ACM Press, New York, New York, USA (2015). ISBN 9781450327589. doi:10.1145/2723372.2749431.

Chung, E. S., Davis, J. D., et al. LINQits: Big Data on Little Clients. In *Proceedings of the 40th Annual International Symposium on Computer Architecture - ISCA '13*, pages 261–272 (2013). ISBN 9781450320795. ISSN 01635964. doi:10.1145/2485922.2485945.

Cohen, J., Dolan, B., et al. MAD Skills : New Analysis Practices for Big Data. *Proceedings of the VLDB Endowmen*, 2(2):1481–1492 (2009). ISSN 2150-8097. doi:10.14778/1687553.1687576.

Cone, M. *The Markdown Guide*. Mark Cone, International e-Book (2018).

Daas, P. J., Puts, M. J., et al. Big data as a source for official statistics. *Journal of Official Statistics*, 31(2):249–262 (2015). ISSN 0282423X. doi:10.1515/JOS-2015-0016.

Dadzie, A. S., Iria, J., et al. The XMediaBox: Sensemaking through the use of knowledge lenses. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5554 LNCS, pages 811–815. Springer-Verlag (2009). ISBN 3642021204. ISSN 03029743. doi:10.1007/978-3-642-02121-3_61.

Davenport, T. H. The Human Side of Big Data and High-Performance Analytics. Technical Report August, Institute for Analytics (2012).

De Mauro, A., Greco, M., et al. A formal definition of Big Data based on its essential features (2016a). doi:10.1108/LR-06-2015-0061.

De Mauro, A., Greco, M., et al. Beyond Data Scientists: a Review of Big Data Skills and Job Families. In *International Forum on Knowledge Asset Dynamics 2016*, pages 1844–1857 (2016b). ISBN 9788896687093.

Dijcks, J. Oracle: Big Data for the Enterprise. Technical report, Oracle Corporation, Redwood Shores, CA (2013).

Dongare, D. and Kadroli, V. Panda: Public auditing for shared data with efficient user revocation in the cloud. In *Proceedings of 2016 Online International Conference on Green Engineering and Technologies, IC-GET 2016*, volume 8, pages 92–106 (2017). ISBN 9781509045563. ISSN 0743166X. doi:10.1109/GET.2016.7916617.

Dryden, I. L. and Hodge, D. J. Journeys in big data statistics. *Statistics and Probability Letters*, 136:121–125 (2018). ISSN 01677152. doi:10.1016/j.spl.2018.02.013.

Dumbill, E. Making Sense of Big Data. *Big Data*, 1(1):1–2 (2013). ISSN 2167-6461. doi:10.1089/big.2012.1503.

Editorial. What can we learn from the Facebook–Cambridge Analytica scandal? *Significance, The Royal Statistical Society*, page 4 (2018).

Ekbia, H., Mattioli, M., et al. Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, 66(8):1523–1545 (2015). ISSN 23301643. doi:10.1002/asi.23294.

Elgendy, N. and Elragal, A. Big Data Analytics in Support of the Decision Making Process. *Procedia Computer Science*, 100:1071–1084 (2016). ISSN 1877-0509. doi: 10.1016/J.PROCS.2016.09.251.

Eppler, M. Visual Literacy (2020).
URL: *https://www.visual-literacy.org/*

EPSRC. Clarifications of EPSRC expectations on research data management . Technical Report October, EPSRC (2014).

Fisher, D., DeLine, R., et al. Interactions with big data analytics. *Interactions*, 19(3):50 (2012). ISSN 10725520. doi:10.1145/2168931.2168943.

Fosso Wamba, S., Akter, S., et al. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165:234–246 (2015). ISSN 0925-5273. doi:10.1016/J.IJPE.2014.12.031.

Fox, C. and Levitin, A. The notion of data and its quality dimensions. *Information Processing and Management*, 30(1):9–19 (1994). ISSN 0306-4573. doi:10.1016/0306-4573(94)90020-5.

Gantz, J. and Reinsel, D. THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digi tal Shadows, and Biggest Growth in the Far East. *Idc*, 2007(December 2012):1–16 (2012).

Gentleman, R. and Temple Lang, D. Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics*, 16(1):1–23 (2007). ISSN 1061-8600. doi:10.1198/106186007X178663.

GnuPG. The GNU Privacy Guard (2013).
URL: *https://gnupg.org/*

Guarino, A. Digital Forensics as a Big Data Challenge. In *ISSE 2013 Securing Electronic Business Processes*, pages 197–203. Springer Vieweg, Wiesbaden (2013). ISBN 978-3-658-03370-5. doi:10.1007/978-3-658-03371-2_17.

Havens, T. C., Bezdek, J. C., et al. Fuzzy c-Means algorithms for very large data. *IEEE Transactions on Fuzzy Systems*, 20(6):1130–1146 (2012). ISSN 10636706. doi:10.1109/TFUZZ.2012.2201485.

Hill, S. and Scott, R. Developing an Approach to Harvesting, Cleaning, and Analyzing Data from Twitter Using R. *Information Systems Education Journal*, 15(3):42–54 (2017).

Hochachka, W. M., Fink, D., et al. Data-intensive science applied to broad-scale citizen science (2012). doi:10.1016/j.tree.2011.11.006.
URL: *https://www.sciencedirect.com/science/article/pii/S0169534711003296*

House of Lords Select Comittee. Artificial Intelligence AI in the UK : ready , willing and able? Technical Report March, House of Lords, London, UK (2018). doi:10.1145/3173574.3174014.

IBM. Big Data Analytics | IBM Analytics (2018).
URL: *https://www.ibm.com/analytics/hadoop/big-data-analytics*

IEEE. About Sharing Your Data and Code - IEEE Author Center (2018).
URL: *http://ieeeauthorcenter.ieee.org/create-your-ieee-article/use-authoring-tools-and-ieee-article-templates/about-managing-your-data/*

Intel IT Center. Intel's 2014 IT Manager Survey on How Organizations Are Using Big data. Technical Report August 2012, Intel IT Center (2014). doi:10.1007/978-3-319-10665-6.

Jacobs, A. The Pathologies of Big Data. *Queue*, 7(6):10 (2009). ISSN 15427730. doi:10.1145/1563821.1563874.

Jain, A. K., Murty, M. N., et al. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323 (1999). ISSN 03600300. doi:10.1145/331499.331504.

Kandel, S., Paepcke, A., et al. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *ACM Human Factors in Computing Systems (CHI)*, page 10. Vancouver, BC, Canada. (2011).

Kelling, S., Hochachka, W. M., et al. Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience*, 59(7):613–620 (2009). ISSN 0006-3568. doi:10.1525/bio.2009.59.7.12.

Kitchin, R. and Lauriault, T. P. Small data in the era of big data. *GeoJournal*, 80(4):463–475 (2015). ISSN 03432521. doi:10.1007/s10708-014-9601-7.

Kitchin, R., Lauriault, T. P., et al. Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. *Regional Studies, Regional Science*, 2(1):6–28 (2015). ISSN 2168-1376. doi:10.1080/21681376.2014.983149.

Knuth, D. E. Literate Programming. *The Computer Journal*, 27(2):97–111 (1984). ISSN 0010-4620. doi:10.1093/comjnl/27.2.97.

Kulkarni, S. and Takawale, N. To study the application of Data Visualization and Analysis tools. *International Research Journal of Multidisciplinary Studies*, 1(5) (2015). ISSN 2454-8499.

Latex Team. Latex Typsetting Software (2020).
URL: *https://www.latex-project.org/*

Lawrence, B., Jones, C., et al. Citation and peer review of data: moving towards formal data publication. *International Journal of Digital Curation*, 6(2):4–37 (2011). ISSN 1746-8256. doi:10.2218/ijdc.v6i2.205.

Layton, R. *Learning Data Mining with Python*. Technical University of Denmark (2015). ISBN 9781784396053.

Lazer, D., Kennedy, R., et al. The parable of google flu: Traps in big data analysis. *Science*, 343(6176):1203–1205 (2014). ISSN 10959203. doi:10.1126/science.1248506.

Leek, J. Six Types Of Analyses Every Data Scientist Should Know - Data Scientist Insights (2013).
URL: *https://datascientistinsights.com/2013/01/29/six-types-of-analyses-every-data-scientist-should-know/*

Lewis, K. P., Vander Wal, E., et al. Wildlife biology, big data, and reproducible research. *Wildlife Society Bulletin*, 42(1):172–179 (2018). ISSN 19385463. doi:10.1002/wsb.847.

Lewis, S. C., Zamith, R., et al. Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting & Electronic Media*, 57(1):34–52 (2013). ISSN 0883-8151. doi:10.1080/08838151.2012.761702.

LIGO Scientific Collaboration. LIGO Open Science Center release of GW150914 (2016). doi:10.7935/K5MW2F23.
URL: *https://losc.ligo.org/events/GW150914/*

Lovelace, R. and Cheshire, J. Introduction to visualising spatial data in R. Technical report, Leeds University, Leeds (2017).

Macfarlane, J. Pandoc User's Guide. Technical report, University of California, Berkeley, CA (2017).

Madin, J., Bowers, S., et al. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3 SPEC. ISS.):279–296 (2007). ISSN 15749541. doi:10.1016/j.ecoinf.2007.05.004.

Manovich, L. Trending: The Promises and the Challenges of Big Social Data. In Gold, M. K., editor, *Debates in the Digital Humanities*, pages 1–10. Univ Of Minnesota Press, Minneapolis (2011). ISBN 9780816677948.

Manyika, J., Chui, M., et al. Big data: The next frontier for innovation, competition, and productivity. Technical Report June, McKinsey Global Institute (2011). doi:10.1080/01443610903114527.

Mayer-Schonberger, V. and Cukier, K. *Big data: a revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt (2013). ISBN 9781848547926.

Microsoft. The Big Bang: How the Big Data Explosion Is Changing the World (2013).
URL: *https://news.microsoft.com/2013/02/11/the-big-bang-how-the-big-data-explosion-is-changing-the-world/*

Microsoft. Digitally sign your macro project - Office Support (2018).
URL: *https://support.office.com/en-us/article/Digitally-sign-your-macro-project-956E9CC8-BBF6-4365-8BFA-98505ECD1C01#ID0EAABAAA=Newer_Versions*

Microsoft. Excel specifications and limits (2020).
URL: *https://support.office.com/en-US/article/Excel-specifications-and-limits-CA36E2DC-1F09-4620-B726-67C00B05040F*

Mohan, K. and Pearl, J. Graphical Models for Processing Missing Data. *arXiv* (2018).

National Biodiversity Network. Mapping - Watsonian Vice Counties - National Biodiversity Network (2018).
URL: *https://nbn.org.uk/tools-and-resources/nbn-toolbox/watsonian-vice-county-boundaries/*

Natural History Museum. Search for a UK species - Natural History Museum (2017).
URL: *http://www.nhm.ac.uk/our-science/data/uk-species/species/index.html*

OED Online. Oxford English Dictionary (2020).
URL: *https://www.oed.com/view/Entry/296948*

Ognyanova, K. Network Analysis and Visualization with R and igraph (2016).
URL: *http://www.kateto.net/wp-content/uploads/2016/01/NetSciX_2016_Workshop.pdf*

Palmer, P., Henshaw, M., et al. A Modular Task Orientated Approach for the Analysis of Large Datasets. *OSF 10.31219/osf.io/ys2vw* (2019). doi:10.31219/osf.io/ys2vw.

Palmer, P. J., Williams, D. J., et al. Observations and models of technology trends within the electronics industry. *Engineering Science and Education Journal*, 8(5):233–240 (1999). ISSN 0963-7346.

Parsons, M. A., Duerr, R., et al. Data citation and peer review (2010). doi:10.1029/2010EO340001.
URL: *http://doi.wiley.com/10.1029/2010EO340001*

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710 (1995).

Pearl, J. The new science of cause and effect, with reflections on data science and artificial intelligence. In *2019 IEEE International Conference on Big Data (Big*, page 4. Institute of Electrical and Electronics Engineers (IEEE) (2020). doi:10.1109/bigdata47090.2019.9005644.

Pearl, J., Li, A., et al. *Causal inference in statistics: A Primer* (2016). ISBN 1935-7516. doi:10.1214/09-SS057.

Pearl, J. and Mackenzie, D. *The book of why : the new science of cause and effect*. Allen Lane (2018). ISBN 978-0241242636.

Pebesma, E. Simple features for R: Standardized support for spatial vector data. *R Journal*, 10(1):439–446 (2018). ISSN 20734859. doi:10.32614/rj-2018-009.

Piwek, L. Tufte in R (2015).
URL: *http://motioninsocial.com/tufte/*

Plackett, R. L. Karl Pearson and the Chi-Squared Test. *International Statistical Review / Revue Internationale de Statistique*, 51(1):59 (1983). ISSN 03067734. doi:10.2307/1402731.

Quarteroni, A. The role of statistics in the era of big data: A computational scientist' perspective. *Statistics and Probability Letters*, 136:63–67 (2018). ISSN 01677152. doi:10.1016/j.spl.2018.02.047.

R Core Team. Data analysis using data.table (2017).
URL: *https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html*

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing (2020).
URL: *https://www.r-project.org/*

Ramos, A. L., Ferreira, J. V., et al. Revisiting the similar process to engineer the contemporary systems. *Journal of Systems Science and Systems Engineering*, 19(3):321 <last_page> 350 (2010). ISSN 1004-3756. doi:10.1007/s11518-010-5144-8.

Rodríguez-Mazahua, L., Rodríguez-Enríquez, C. A., et al. A general perspective of Big Data: applications, tools, challenges and trends. *Journal of Supercomputing*, 72(8):3073–3113 (2016). ISSN 15730484. doi:10.1007/s11227-015-1501-1.

Royal Society. *Machine learning : the power and promise of computers that learn by example*, volume 66 (2017). ISBN 9781782522591.

RStudio Team. RStudio (2016).
URL: *http://www.rstudio.com/*

Russom, P. Big data analytics. *TDWI best practices report, fourth quarter*, 19(4):1–34 (2011). ISSN 97836423.

Sagiroglu, S. and Sinanc, D. Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 42–47. IEEE (2013). ISBN 978-1-4673-6404-1. ISSN 9781467364034. doi:10.1109/CTS.2013.6567202.

Saunders, M. N. K., Lewis, P., et al. *Research methods for business students*. Pearson Education Ltd, Harlow, UK, seventh edition (2016). ISBN 9781292016627.

Schneble, C. O., Elger, B. S., et al. The Cambridge Analytica affair and Internet-mediated research. *EMBO reports*, page e46579 (2018). ISSN 1469-221X. doi:10.15252/embr.201846579.

Shneiderman, B. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343 (1996). ISBN 0-8186-7508-X. ISSN 1049-2615. doi:10.1109/VL.1996.545307.

Shneiderman, B. Extreme visualization. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data - SIGMOD '08*, pages 3–12 (2008). ISBN 9781605581026. ISSN 07308078. doi:10.1145/1376616.1376618.

Sivarajah, U., Kamal, M. M., et al. Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70(August):263–286 (2017). ISSN 01482963. doi:10.1016/j.jbusres.2016.08.001.

Slater, S., Joksimovic, S., et al. Tools for Educational Data Mining: A Review. *Journal of Educational and Behavioral Statistics*, 42(1):85–106 (2016). ISSN 1076-9986. doi:10.3102/1076998616666808.

Stodden, V., Leisch, F., et al. *Implementing reproducible research*. CRC Press/Taylor and Francis (2014). ISBN 1466561599.

Suthaharan, S. Big Data Classification : Problems and Challenges in Network Intrusion Prediction with Machine Learning. *Performance Evaluation Review*, 41(4):70–73 (2014). ISSN 0163-5999. doi:10.1145/2627534.2627557.

Tenopir, C., Allard, S., et al. Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6(6):e21101 (2011). ISSN 19326203. doi:10.1371/journal.pone.0021101.

The CA / Browser Forum. Baseline Requirements for the Issuance and Management of Publicly-Trusted Certificates, v.1.0. Technical report, CA / Browser Forum (2011).

Tsichritzis, Dionysios C and Lochovsky, F. H. *Data models*. Prentice Hall Professional Technical Reference (1982).

Tufte, E. R. *The Visual Display of Quantitative Information*. Graphics Press, Nuneaton, UK, 2nd edition (2001). ISBN 978-1930824133.

Tumminello, M., Micciché, S., et al. Statistically validated networks in bipartite complex systems. *PLoS ONE*, 6(3):e17994 (2011). ISSN 19326203. doi:10.1371/journal.pone.0017994.

Tyner, S., Briatte, F., et al. Network Visualization with ggplot2. *The R Journal*, 9(1):27–59 (2017). ISSN 2073-4859.

UK Government. The Natural Environment and Rural Communities Act 2006. *Environmental Law Review*, 8(4):292–298 (2006). ISSN 1461-4529. doi:10.1350/enlr.2006.8.4.292.

Vetrò, A., Canova, L., et al. Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, 33(2):325–337 (2016). ISSN 0740624X. doi:10.1016/j.giq.2016.02.001.

Wang, L., Wang, G., et al. Big Data and Visualization: Methods, Challenges and Technology Progress. *Digital Technologies*, 1(1):33–38 (2015). doi:10.12691/dt-1-1-7.

Ward, J. S. and Barker, A. Undefined By Data: A Survey of Big Data Definitions. *arXiv* (2013). ISSN 00010782. doi:10.1145/2699414.

Wickham, H. Tidy Data. *Journal of Statistical Software*, 59(10):1–22 (2014). ISSN 1548-7660.

Wickham, H. *Advanced R, Second Edition (Chapman & Hall/CRC The R Series)*. London, UK, 2 edition (2019). ISBN 0815384572.

Wickham, H. and Bryan, J. *R Packages*. O'Reilly Media, Boston, MA (2015). ISBN 978-1491910597.

Wickham, H. and Grolemund, G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media (2016). ISBN 9781491910399. doi:10.1145/3138860.3138865.

Wilkinson, L. *The grammar of graphics* (2010). doi:10.1002/wics.118.

Williams, C. K. and Wynn, D. E. A critical realist script for creative theorising in information systems. *European Journal of Information Systems*, 27(3):315–325 (2018). ISSN 14769344. doi:10.1080/0960085X.2018.1435231.

Wong, L. Big data and a bewildered lay analyst. *Statistics and Probability Letters*, 136:73–77 (2018). ISSN 01677152. doi:10.1016/j.spl.2018.02.033.

Wu, X., Zhu, X., et al. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):97–107 (2014). ISSN 10414347. doi:10.1109/TKDE.2013.109.

Xie, Y. *Bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman and Hall/CRC, London, UK (2017). ISBN 9781138700109.

Yakout, M., Elmagarmid, A. K., et al. Guided data repair. In *Proceedings of the VLDB Endowment*, volume 4, pages 279–289. Association for Computing Machinery (2011). ISSN 21508097. doi:10.14778/1952376.1952378.

Yu, S., Liu, M., et al. Networking for Big Data: A Survey. *IEEE Communications Surveys & Tutorials*, 19(1):531–549 (2017a). ISSN 1553-877X. doi:10.1109/COMST.2016.2610963.

Yu, W., Carrasco Kind, M., et al. Vizic: A Jupyter-based interactive visualization tool for astronomical catalogs. *Astronomy and Computing*, 20:128–139 (2017b). ISSN 22131337. doi:10.1016/j.ascom.2017.06.004.

Zanin, M., Papo, D., et al. Combining complex networks and data mining: Why and how (2016). doi:10.1016/j.physrep.2016.04.005.

Zhao, Y. R and Data Mining: Examples and Case Studies. Technical report, RDataMining (2015). doi:10.1016/B978-0-12-396963-7.00001-5.

Appendix A

Literature Review Method

The method used in this work was inspired by Sivarajah et al. (2017) in their review of 'Big Data challenges and analytic methods' which had to address the same problems of filtering results. Sivarajah sought to align their approach with to a Systematic Literature Review (SLR), by using three well-defined phases: Planning; Searching; Review and synthesis. Their description of the search process requires a review and conceptualisation of results as an essentially linear process, which makes their presentation of the outcomes flow naturally from the search results. This work develops the process by regarding the selection and definitions of keywords as an iterative process that is developed by interpretation of the results using a systems style approach (Ramos et al., 2010).

The two primary search tools used in this literature review were:

Google Scholar `https://scholar.google.co.uk`;

Web of Science `https://apps.webofknowledge.com`.

The effectiveness of these search tools in part due to comprehensive nature of the results returned, and in part due to the open access of many academic publications allowing for full-text to be downloaded for reading. However, simplistic search strategies involving the term 'Big Data' return 4.9 million results on Google Scholar and 57,000 results for Web of Science. It is ironic that an exploration of this domain rapidly runs into problems relating to the volume of information available. Without *a priori* knowledge of the domain of interest

and keywords used by practitioners, the refinement of results by using more complex search terms is problematic if inappropriate selection and rejection of references are to be avoided.

Both search tools are in common use within the academic community and both permit the saving and export of search results into third party referencing tools such as:

Mendeley `https://www.mendeley.com`

It should be noted that both these tools limit downloads. Google forces the user to select articles for inclusion into a personal library, and then allows downloads in groups of twenty items, imposing a manual element to the generation of a large corpus, which itself comprises multiple files. Web of Knowledge allows a greater number of downloads in a single transaction, but still requires manual intervention. Circumventing these controls would be a breach of the licence terms.

The literature search process was broken down into a series of sub-tasks conceptually represented in Figure A.1 and started with *Web of Science*, in part because of the ease with which more sophisticated search queries may be applied. This process is similar to that adopted by Rodriguez *et al* Rodríguez-Mazahua et al. (2016) was used to extract the list of keywords given in Table A.1 from corpus of references using the following method:

- This iterative process started from an initial set of 'Big Data' papers that the author regarded as 'interesting' and 'on-topic'. Although this appears to be a vague and serendipitous way to start a research process, it builds on the strengths of digital search strategies to apply filtering on a large pool of results. An objective at this stage is to find unexpected keywords 'hidden' within the discovered sources.
- A selection of papers referring to 'Big Data' were found using Google Scholar and saved to generate an initial corpus to seed the second step. A similar corpus was saved using Web of Science.
- The recurring words were extracted from the title of papers in the corpus using an R script written by the author and summarized in Table A.1. [1]

Records were not kept of papers that were not considered relevant due to the potential size of this list. The following general policies were applied to sources selected for a more

[1] The R markdown document is included as Appendix . The code was written to work with any corpus of documents extracted from Google Scholar or Web of Science.

word
big
data
mining
bibliometrics
research
cleaning
knowledge
analysis
applications
discovery
methods
techniques
information
taxonomy
system
privacy
visualization

Table A.1: Initial research keywords

detailed interpretation. Although these policies could be viewed as arbitrary, it should be remembered that they were applied to search results that sometimes returned millions of raw hits, and served the purpose to reduce search results to a manageable number.

For each case the search results were quickly assessed using the following process:

Skim title Paper titles were read and a view formed on their relevance. If too few appeared relevant the search was abandoned and a fresh permutation of keywords used. Further refinements by article type, publication, relevance, number of citations were used to bring the list down to something that could be scanned by eye to select an initial collection for review.

Read abstract The abstract was only read when papers seemed of interest. Most were discarded on the basis of the abstract.

Interpret body Where a source was of sufficient interest the main body of content was reviewed. *Papers were rejected if the body could not be accessed.*

Critque knowledge Papers passing all the above stages were imported into Mendeley for an in-depth review. A proportion of these were considered relevant for this review.

Suggest keywords All the preceding stages contributed to the development of keywords for searches.

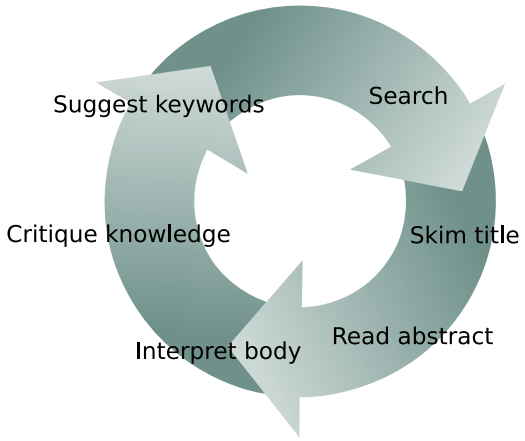


Figure A.1: Conceptual literature search process.

Various factors that may have an impact on the perceived relevance of papers are summarised in Table A.2. The main point here is to stress the importance of title and abstract in the paper selection process. While paper readability is essential, there are several preceding stages in the search process that might eliminate otherwise relevant sources.

Pros	**Cons**
Informative title aligns well with subject.	Obtuse title.
Document Open access or freely available	Full paper not easily available.
Abstract concise and describes key content.	Abstract not easy to contextually interpret.
Single column format available for speed reading online	Paper layout hard to read on screen.

Table A.2: Pros and Cons of literature review methodology

Academic Peer reviewed academic publications were strongly preferred as a source for this review. Newspapers, industry magazines, opinion pieces and blogs were also reviewed where they contained relevant information. Where primary academic sources were cited within such non-academic documents, these were reviewed independently and treated as the primary source.

Access Open Access documents were strongly preferred as sources. Sources that provided only summary access were not included. While this may exclude some useful older references or books, it was considered likely that such sources would be visible because they are repeatedly cited within the general literature.

Bibliographic Collation Every reference was collated within the Mendeley reference tool and a full entry created that included either a PDF document or a permanent URL. For web sources a PDF was created at the time of access and stored with the reference.

Language Only papers written in English were considered for inclusion. (Because the author only speaks English fluently.)

Relevance References relating to GRID computing and the huge datasets generated by parts of the physics community were eliminated as not relevant to this work. Most references to the biosciences, especially genomics were also considered not relevant. Some, however, did consider desktop computing so were evaluated.

A.1 Comments On Methodology

This is a pragmatic methodology based on that used by Sivarajah et al. (2017) in their review of 'Big Data challenges and analytic methods', following a formal definition of process similar to a systematic literature review, but focussing on on-line searches.

The bias introduced by the rejection of sources that were not accessible online is potentially mitigated by the observation that key sources may come to attention through frequent citation by multiple authors in papers selected for in depth review. This was found especially to be the case for books such as: Bowker, G. C. et al. (2013) 'Raw data' is an oxymoron; Bowker et al. (2013), which were found by following references. As an aside, examples were found where the reference citation context did not match the source content, so such sources were discarded. Even with these problems, the principle merit is that the process is achievable within a reasonable time frame and the corpus of material for deeper review is generated by a systematic process. Other authors such as Rodríguez-Mazahua et al. (2016); De Mauro et al. (2016a) also described broadly similar strategies to manage the number of results that arise from the use of modern search engines. As a related issue, it is also becoming common to find review papers citing hundreds of references, making follow up of

all secondary sources too time-consuming to be practical (Boccaletti et al., 2014; Yu et al., 2017a; Zanin et al., 2016). A principal concern is that important sources will be missed in the overall clutter, although this has been addressed to some extent by adopting the circular process in Figure A.1 which incorporates iterative updating of the search keywords on the basis of information discovered.

	Exploratory	Programmatic
Technology	Hardware and physical systems.	Software and interfaces.
Normative	The ' How to' of analysis.	Describing programmatic techniques for visualising data.
Informative	Exploring the concepts of data exploration and visualisation.	Models of programming for knowledge extraction.
Impact	Discussing the societal and commercial impacts.	Describing predictive models of societal change.

Table A.3: Big Data themes and related topics in existing literature. After De Mauro et al. (2016a)

Table A.3 captures as a matrix a possible high level classification method for papers. The author notes that many reviews are 'informative' in nature (as in Wu *et al* above) and may not be strongly coupled to current 'normative' sources given their potentially deep technical nature. This may lead to a lag in the appreciation of the capability of desktop and mobile computing. For example, Chung *et al* present specialised SoC techniques that have the potential to migrate Big Data type applications on to the ARM architecture used on smartphones Chung et al. (2013).

Appendix B

Stakeholder Interviews

B.1 Introduction

Interviews were conducted in semiformal sessions at the stakeholders place of work, and it was made clear when booking the interview, a time slot of about an hour would be required. Duration was only extended at the interviewees discretion and care taken to respect any time constraints. That said, an overriding common theme emerged of data accumulating faster than stakeholders ability to analyse with the tools to hand. As a consequence, there was an unanticipated enthusiasm for participation in this research which was seen as potentially helping resolve a widely recognised analytical shortfall.

The preceding words have been carefully framed to externalise rather than personalise the analytical challenges discussed in the interviews. This is a deliberate interpretation intended to emphasise the challenges brought about through increasing data size, rather than implying any shortcomings in the ability of the stakeholders. The current approach is to squeeze evermore sophisticated analyses into existing tools by spending more time hand annotating the results. It is truly impressive what is achieved, but everyone believes there has to be a better way forward.

A synthesised summary of the interview results are presented in the following sections.

B.2 Structured Questions

The questions were focussed on building an understanding of the whys and whats of data collection, followed by questions of analytical techniques used. This emphasis naturally follows from the choice in methodology to regard **antecedent data** as an area of research importance. There was always a risk that this approach to questioning would find that all data are collected and reported under controlled conditions, or at least, that stakeholders would give that impression, undermining the concept of antecedent data. However, while the academic terminology adopted here was not used by stakeholders, the underlying concepts proved to be valid representations of real world issues.

The questions used in the meeting pro-forma are presented in Table B.1 along with a brief narrative explanation of the intention behind each question. The formulation of these questions was arrived at through a process of abduction based in part of the author's knowledge of the domain and partly through very informal pre-interview discussions. Despite such *a priori* knowledge, the candour of results were surprising to the author.

Question	**Intention**
What data do you collect?	A
Why do you collect data?	A
How would you like to summarise data?	A
What tools do you currently use?	A
Have you seen someone else address a problem that you would also like to solve?	A
Would you like to publish in peer reviewed journals?	A
Do you have a specific idea that you would like me to look at for you?	A

Table B.1: Summary interview questions

B.3 Qualitative Data Analysis

- Thirteen interviews were made with end users and stakeholders.
- A qualitative data analysis was performed on the notes using Latex ulqda package
- Summary QDA to be produced using R Template.

Appendix C

Motivational Example Template

A working version of the motivational example template is presented in this unlisted YouTube video which may be viewed from this link:
`https://youtu.be/zBQ_ypVYDWg`

The template is split into two parts. The first extracts data from a disjointed set of Excel spreadsheet. This is a slow process as it has to search through all the files and sub-sheets. Synonyms for places and observations are replaced with current preferred names. Finally, data *s.s* are saved in several formats to allow import into other software.

The second template produces performance indicators from data *s.s*. The pair of templates may be re-run every time the source raw-data are updated.

Appendix D

Interview notes

D.1 Rutland Water Stakeholders

Summary of Rutland Water Nature Reserve
End-User Meeting 2018-11-09

Paul J Palmer

09 November 2018

1 Discussion

1.1 Purpose

- To understand the type and extent of data collated around the Rutland Water environs.
- To understand issue around the use of that data.
- To understand how improved analytic tools might be of use.

All of these questions are associated with analysis of biodiversity data.

1.2 Present

- Manager.
- Senior Reserves Officer.
- PhD Researcher, Loughborough University.

1.3 Background

Rutland Water is a large man-made reservoir created in 1975. Rutland Water Nature Reserve is unique in that it was declared a reserve before it existed. The wildlife potential of the proposed reservoir was recognised as early as 1969; reserve boundaries and the construction of lagoons were formulated in 1972 and in 1975 the Trust signed a management agreement with the Anglian Water Authority. In 2002 the areas managed by the Wildlife Trust were increased to include Barnsdale, Armley, and Hambleton Woods and Berrybutts Spinney `https://www.lrwt.org.uk/nature-reserves/rutland-water/`.

The following designations relate to the quality and management of the site:

- Local Wildlife Site
- NATURE CONSERVATION REVIEW
- Ramsar Site
- Site of Special Scientific Interest

- Special Protection Area
- Water Framework Directive (WFD)

2 Questions

2.1 What data do you collect?

Data collection is not well-organised; in reality every volunteer recorder is different and uses as an individual methodology. Many recorders working with specialist areas (such as entomology) collate information and submit directly to county or national schemes. Not all such data flows through the reserve office. In theory all such data should end up as on the ORCA database operated by LRERC where it may be accessed through a closed portal, but in practice, not all records make it to ORCA.

what data!poor organisation, data!multiple!systems, data!multiple!methodolgies, data!multiple! data flows, data!multiple!recording schemes, data!no stanaard data flow, data!no standard destination

Birds records are often directly submitted to the reserve office in a variety of format including paper. Paper copies of the reserve visitor sightings daybook back to the 1970's.

Breeding birds presence survey records are available. Each year the survey produces around 1200 sheets of marked up high resolution maps, with breeding bird observations. However, the digitisation process loses metadata resulting in very simplistic summaries of the field data.

what!multiple!formats, formats!paper, formats!electronic

A custom App is currently under consideration as a means to digitise data at the point of observation.

what!volume, data!metadata!loss

what!aspirations!custom recording App

2.2 Why do you collect data?

Wildfowl counts (WeBS) and water level records are mandatory. The collation of all other records are part of the job.

why!mandatory, part of the job

2.3 How would you like to summarise data?

Linking observed data to reserve management goals is a current aspiration. For example, the lagoon water levels is managed, but what water level regime results in the highest number of breeding birds?

how!aspirations!analysis!reserve management

Some work is currently active in this area: Dr. Sarah Johnson NERC Earth Observation project for wetlands is on secondment for six months from Leicester University looking related environmental issues.

how!current!satellite data

2.4 What tools do you currently use?

No specific tools are in current use.

tools!analysis!no common tools

2.5 Have you seen someone else address a problem that you would also like to solve?

Looking at the effects of climate change on specific frequently seen species: Mallard, Pochard. Egrets.

aspirations!measure climate change

2.6 Would you like to publish in peer reviewed journals?

Yes. Time stops publishing. Aspire to publish Nature! *aspirations!publish in journals*

There is a real gap in using the data that we collate and any liaison with a University would be a big help in closing that gap. *aspirations!use data that is collected*

2.7 Do you have a specific idea that you would like me to look at for you?

Combining data from many sources. Satellite, weather, land management. It is very difficult for us to analyse data, but ideally the ability to combine data from many sources and overlay that on reserve management maps. *aspirations!combining data*

A best practice of analysis that results in a paper would be an encouragement to both staff and volunteers. *aspirations!write best practice paper*

3 Concluding Notes

Although the data size is small, the variety is high which presents a Big Data type challenge. There is no formal control of the data collation process, although all biodiversity records should ultimately end up in ORCA and thus be accessible for analysis. *data!size!small; data!variety!high* *data!multiple! data flows, data!ORCA*

A meeting which will discuss the link to ORCA (Andy Lear) has been arranged 23/11/2018.

No specific tools are in use. A slight bias against the use of Open Source was noted. The current failings of the popular proprietary 'Recorder' package were also noted.[1] *analysis!no common tools*

[1] Recorder has its origins as MS-DOS software and was adopted by many local environmental records centres. There is currently limited support for the application.

D.2 LRWT Stakeholders

Summary of LRWT End-User Meeting
2018-11-23

Paul J Palmer

23 November 2018

1 Discussion

1.1 Purpose

- To understand the type and extent of data collated around the Rutland Water environs.
- To understand issues using data.
- To understand how improved analytic tools might be of use.

All of these questions are associated with analysis of biodiversity data.

1.2 Present

- Conservation Officer. LRWT.
- Conservation Officer. LRWT
- PhD Researcher, Loughborough University. [Paul J. Palmer]

1.3 Background

Leicestershire and Rutland Wildlife Trust was founded in 1956 by a small group of naturalists and was formerly known as the Leicestershire and Rutland Trust for Nature Conservation. It is a registered charity concerned with all aspects of nature conservation. The Trust has a professional team of 25 staff and more than 500 active volunteers. It manages 35 nature reserves.

URL: `https://www.lrwt.org.uk/`.

2 Questions

2.1 What data do you collect?

Leicestershire biodiversity data. The LRWT Record system shows over 629,000 records. However, data management not very well integrated and the central system does not link well to Rutland Water. Data very skewed from Nature Reserves and is not comprehensive.

what!data!leicestershire biodiversity, data!size, data!poor integration

There is now a data exchange with ORCA [1] so the records are held in two places for security . There is a concern that ORCA and LRERC are very dependent on county council budget.

what!data! permanence

In recent years fewer people submit records. It is presumed that the growth of on-line recording schemes is the main reason for this. Also not all the county recorders [2] pass data to the LRWT. Periodically data are incorporated from the NatureSpot and iRecord online systems. The trust has digitised much of the historic paper data for all reserves except Rutland Water. There is no facility for duplicate checking, so the magnitude of this problem has not been accurately quantified.

what!data!reduction

what!data!synchronisation, what!data!digitisation

what!data!error checking!none

2.2 Why do you collect data?

Data are collected to support biodiversity management at county and local levels. At county level it informs policy relating to meeting conservation targets [3]. At a local level it monitors the effectiveness of management of existing holdings and identifies potential future areas of interest.

why!management ! monitor, why!management!identify future holdings

2.3 How would you like to summarise data?

The LRWT would like to be able to report its environmental holdings to support targeted management and publicity. For example, understanding what taxa are present on its reserves would enable targeted purchasing of new localities to maximise the holdings. Publicising the proportion of taxon protected would be used for promotional purposes.

how!data! Reporting taxa, how!Management!Targeted purchase

A narrative report for this purpose was hand assembled from data. Its production was very time-consuming.

how!Analysis!manual!Slow

[1] ORCA is the system used by the Leicestershire and Rutland Environmental Records Centre.

[2] Each taxa group has a designated county recorder responsible for managing records.

[3] `http://jncc.defra.gov.uk/page-5281`

2.4 What tools do you currently use?

Recorder [4]. This software has its roots in MS-DOS and now has very limited support. The limitations are well understood, but so much effort has been invested in its use, finding a replacement has proven difficult. While not considered ideal, ORCA is the only available alternative.

tools!analysis!Recorder

The principle problem with Recorder is seen as the support for user customisation. The 'atom' of a record is a survey, usually associated with a person, which does not map well to the *ad hoc* nature of most records. Multiple survey templates have resulted in a multiplicity of data fields. For example, there are numerous choices for abundance and location resulting in inconsistent metadata. While the system supports hierarchical naming of localities, this has not been consistently used.

There is a conflict between the needs of data recording and the user interface. This shows up especially in the number of fields that have to be completed with sensible defaults.

tools!Analysis!Conflict ! Recording, tools!Analysis! Conflict! user interface

Reports are created using a time-consuming process to select options. Reports are output as Excel files, but need manually tidying up before use as there are often many unused fields present. In addition, there is no duplicate record removal. Also, Recorder can output several format of Excel - not all work.

tools!Analysis! slow process ! Excel; Analysis! Duplicate removal

R is used on a new project that seeks to analyse data on bat distribution that is being collected by automatic recording equipment.

tools!Analysis! R ! Bats

2.5 What type of analysis would you like to be able to do, but lack resource or expertise?

We would like to quantify proportion of taxon represented in our current reserves when compared to the county as a whole. We want to use this for both PR and to identify habitats that the trust should pruchase.

Analysis! Inventory, Analysis! Support management targets

Give me a list of all the xxxx is the commonest question, but this is really time-consuming to assemble.

Analysis! Give me a list

We have no practical way of eliminating duplicates records.

analysis!Data! Management ! Duplicates

2.6 Would you like to publish in peer reviewed journals?

Yes. Currently we do not do anything with the data.

Analysis ! publish, Data! Archive only

[4] http://jncc.defra.gov.uk/recorder

2.7 Do you have a specific idea that you would like me to look at for you?

Nathalie Cosser is looking at using R to produce maps relating Bat data collected using audio sensors. Any help would be useful.

Analysis! Maps ! bats

A working script to remove duplicate records from reports would be used. There is no problem with installing R on trust computers.

analysis!Data! Duplicates! Remove

3 Concluding Notes

Offer to make available my script that remove duplicates records from Recorder output reports. Meet Nathalie Cosser to discuss use of R.

D.3 LRWT Conservation Committee

Summary of LRWT Conservation Committee 2019-03-11

Paul J Palmer

11th March 20

1 Discussion

1.1 Purpose

- To present species inventory markdown document.

All of these questions are associated with analysis of biodiversity data.

1.2 Present

Conservation Committee Members

PhD Student

1.3 Background

Leicestershire and Rutland Wildlife Trust was founded in 1956 by a small group of naturalists and was formerly known as the Leicestershire and Rutland Trust for Nature Conservation. It is a registered charity concerned with all aspects of nature conservation. The Trust has a professional team of 25 staff and more than 500 active volunteers. It manages 35 nature reserves.

URL: `https://www.lrwt.org.uk/`.

2 Questions

- Asked if the summary includes aggregate species. No it does not, but the Species Dictionary includes aggregates so that this might be included in a future version.
- Can this be adapted to highlight areas outside of reserves that have a high concentration of indicator species? Claire might be able to suggest an indicator list.
- Can I do a workshop to demonstrate the power of the analytical techniques used.
- Concern that open source software might entail opening the access to the source data.
- Great interest in the method used to identify the duplicate records. Once again data cleaning is a major challenge.

question!Anal Aspiration! a of aggregate s

question!anal Aspiration! N

question!Anal Workshop

Tools! Open S Does this mea Data?, Data! Confidentialit

question!Data Cleaning! Du

3 Follow on work

- Confirm validity of assumptions with Ben Devine and Andy Lear.
- All Trust data is being migrated to Orca so the inventory might be applied to the larger data that will result.

question!Data Validity of assumptions

question!Data Analysis! Lar Data! ORCA: Migration of data

1

D.4 LRWT Conservation Committee Species Inventory

A draft wildlife inventory for Leicestershire

Paul J Palmer

30/01/2019

Introduction

This analysis used data provided as a single large excel file by LRWT of all wildlife records with a minimum of filtering. Prior to analysis the data were converted to a standard csv format using the in2csv command from csvkit as the excel file could not be read in a reasonable time due to limitations in the R routines used. Reserve polygons were also provided and used to annotate the records and in or out of reserve polygons. The Reserve polygon boundaries were assumed by the author to be non overlapping - if this is not the case then some records may be counted twice. This process was entirely automated so records near reserves did not count, but those on a boundary line will be counted as in the reserve.

Not all records were complete due to problems with dates and locations. Incomplete dates were left in as they typically indicate the use of date ranges associated with valid observations. Including these data in any plots using time as a variable would require care in interpretation. Incomplete locations were ignored.

The taxon list has been updated against the Natural History Museum Species dictionary supplied 20/02/2019. Each taxon name was matched to the current preferred names. This showed that 351838 entries were duplicates once synonyms had been taken into account. Aggregate names have been ignored in this analysis.

The original data also used reserved characters in names that do not play well with programming languages. Working with these requires extra code to replace them with safe names during analysis to prevent system errors. This is not a criticism, it is an observation that real data should not be treated as if it were compliant with the syntax of programming languages. The resulting errors can be misconstrued as physical memory limitations due to error messages, rather than programming errors.

Results

The summary table 1 presents some basic facts inferred from the data. The proportion of taxon present on reserves, 79 % is an interesting and commendable figure of merit that will stand up to scrutiny, given the automated method used to calculate it. The approximate centres of the reserves are marked on the map shown in figure 1.

Table 2 provides a count of taxon grouped by reserve. As can be seen, not all reserves are equally gifted in terms of taxon present. Finally, the table 3 presents the data by group. This also is derived from the nomenclature used from the NHM Species dictionary so will again stand up to scrutiny.

Table 1: Summary

Description	N
Initial number of rows	629305
Incomplete dates	46121
Dropping bad grid refs leaves	620565
Total number of taxon	6718
Total number duplicate records	351838
Number of taxon in reserves	5327
Percentage of taxon in reserves	79

Probably the greatest area for statistical concern is under recording of areas outside of reserve boundaries artificially increasing the figure of merit.

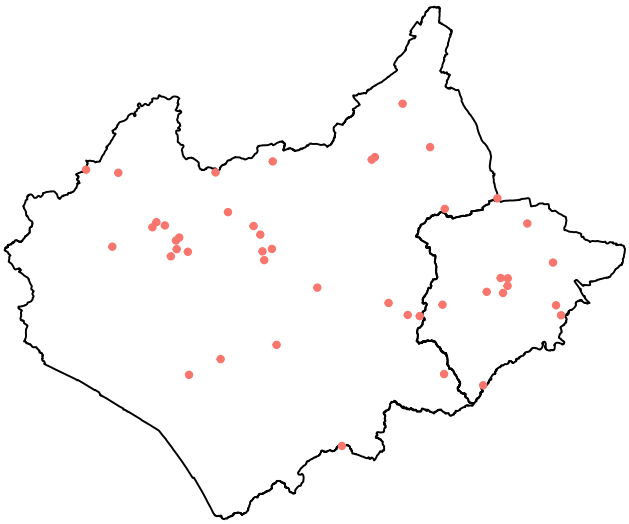


Figure 1: Map of reserve locations

Map of reserves

The map of reserve locations marks the approximate centres of each of the 45 nature reserves.

Table 2: Reserve Summary

Reserve	Count
Altar Stones	209
Armley Wood	32
Barnsdale Wood (West)	183
Bloody Oaks Quarry	416
Charley Wood	574
Charnwood Lodge	1928
Charnwood Lodge (The Chase)	59
Cloud Wood	1165
Coombs Meadow	316
Cossington Meadows	855
Cribb's Meadow	392
Croft Pasture	505
Dimminsdale	601
Fox Wood	3
Great Merrible Wood	264
Hambleton Wood	127
Hambleton Wood (East)	1
Holwell Reserves	791
Holwell Reserves (North Quarry)	18
Kelham Bridge	509
Ketton Fields	110
Ketton Quarry	1383
Launde Big Wood	996
Launde Park Wood	897
Lea Meadows	796
Loughborough Big Meadow	969
Lucas' Marsh	475
Lyddington Meadow	53
Merry's Meadows	300
Mountsorrel and Rothley Marshes	375
Narborough Bog	1333
Prior's Coppice	1009
Rocky Plantation	159
Rutland Water Nature Reserve	1644
Stonesby Quarry	545
Syston Lake	99
The Miles' Piece	1
Tilton Railway Cutting	38
Tom Long's Nature Reserve	35
Ulverscroft	1299
Ulverscroft (Herbert's Meadow)	619
Wanlip	278
Wanlip Rough	84
Wymeswold Meadows	206
Wymondham Rough	565

Table 3: Species summary

Reserve	acrolite (Acari)	alga	amphibian	annelid	bird	bony fish (Actinopterygii)	centipede	chromist	conifer	crustacean	diatom	false scorpion (Pseudoscorpiones)	fern	flatworm (Tricladida)	flowering plant	fraxinifolium	fungoid	fungus	harvestman (Opiliones)	horsetail	insect - beetle (Coleoptera)	insect - bucklisme (Neuroptera)	insect - butterfly	insect - caddis fly (Trichoptera)	insect - dragonfly (Odonata)	insect - flea (Siphonaptera)	insect - hymenopteran	insect - lacewing (Neuroptera)	insect - mayfly (Ephemeroptera)	insect - moth	insect - orthopteran	insect - scorpion fly (Mecoptera)	insect - snakefly (Raphidioptera)	insect - stonefly (Plecoptera)	insect - true bug (Hemiptera)	insect - true fly (Diptera)	jawless fish (Agnatha)	lichen	liverwort	millipede	mollusc	moss	peronosan	reptile	rodifer	slime mould	spider (Araneae)	springtail (Collembola)	starwort	terrestrial mammal	
Altar Stones	0	0	2	0	11	0	0	0	0	0	0	0	2	0	65	0	0	47	0	0	4	0	10	0	0	0	0	0	0	1	0	0	0	0	0	1	0	61	1	0	0	1	0	1	0	0	0	0	2
Armley Wood	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	3
Barnsdale Wood (West)	0	0	0	0	3	0	0	0	3	0	0	0	0	0	140	0	1	27	0	0	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	2
Bloody Oaks Quarry	0	0	0	0	3	0	0	0	0	0	1	0	0	1	160	0	0	24	0	1	9	0	24	0	3	0	2	0	0	151	4	0	0	0	1	25	0	0	0	0	0	0	2	0	0	3	0	0	2
Charley Wood	1	0	1	0	51	0	0	0	3	0	0	0	5	0	142	0	0	30	1	0	11	0	17	0	12	0	9	0	0	139	2	0	0	0	2	26	0	0	21	0	0	66	0	0	0	24	0	0	11
Charnwood Lodge	0	31	7	2	105	1	0	1	14	2	2	1	12	0	314	0	0	135	5	2	262	0	24	13	17	6	16	0	0	421	3	0	0	1	12	118	0	60	33	2	3	113	1	2	8	1	151	1	1	25
Charnwood Lodge (The Chase)	0	0	2	1	7	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	3	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Cloud Wood	2	0	2	0	59	0	0	0	1	2	0	0	10	0	292	0	0	80	1	0	39	0	26	0	7	0	12	0	0	417	5	0	0	0	2	105	0	4	7	3	2	47	0	0	2	21	0	1	16
Coombs Meadow	0	0	1	0	24	0	0	0	0	0	0	0	3	0	158	0	0	1	0	4	1	0	6	0	3	0	0	0	0	103	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	1	0	0	5
Cossington Meadows	1	0	5	6	182	8	0	0	0	5	0	0	0	1	185	0	0	1	0	0	90	0	20	18	22	0	6	0	5	179	5	0	0	0	21	32	0	3	4	0	3	29	0	1	0	0	4	1	1	17
Cribb's Meadow	0	0	4	0	42	1	0	0	0	0	0	0	2	0	198	0	0	3	0	0	10	0	23	0	12	0	5	0	0	64	0	0	0	0	2	9	0	0	0	0	2	0	0	1	0	4	0	2	8
Croft Pasture	0	0	0	0	21	8	0	0	0	0	0	0	2	0	242	0	0	2	0	1	46	0	18	0	2	0	3	0	0	46	1	0	0	0	1	19	0	39	2	0	0	43	0	1	0	0	5	0	0	3
Dinsmiesdale	0	1	3	0	54	1	0	0	1	2	0	0	11	0	208	0	0	90	1	3	22	0	13	1	7	0	1	0	1	100	0	0	0	0	0	9	0	8	7	0	0	44	0	0	0	2	1	0	0	10
Fox Wood	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Great Merrible Wood	0	0	0	0	15	0	0	0	0	0	0	0	11	0	115	0	0	23	1	2	1	0	2	0	2	0	0	0	0	81	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	2	0	0	6
Hambleton Wood	0	0	0	0	44	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	6	0	0	24	0	0	0	0	0	0	8
Hambleton Wood (East)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Holwell Reserves	0	1	2	2	55	0	0	0	3	6	0	0	5	0	220	0	0	0	0	2	4	0	28	0	9	0	4	0	0	282	3	0	0	0	0	61	0	39	3	7	6	23	0	2	0	0	1	0	0	19
Holwell Reserves (North Quarry)	0	0	0	0	2	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Kelham Bridge	0	0	3	0	124	3	0	0	0	0	0	0	0	0	122	0	0	0	0	0	49	0	16	0	11	0	1	0	0	134	2	0	0	0	5	13	0	0	0	0	0	4	0	0	0	0	0	0	22
Ketton Fields	0	0	0	0	2	0	0	0	2	0	0	0	0	0	100	0	0	0	0	0	1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ketton Quarry	0	0	1	0	72	0	1	0	3	2	0	0	13	0	346	0	1	97	2	1	130	0	35	0	15	0	16	0	0	465	6	0	0	0	2	132	0	1	0	3	7	5	0	3	0	2	10	0	0	12
Launde Big Wood	3	0	2	0	47	0	3	0	10	3	0	1	9	0	213	0	2	119	0	3	86	0	26	0	8	0	11	0	0	270	0	0	0	0	0	86	0	1	5	3	6	25	0	0	0	3	37	0	0	14
Launde Park Wood	0	0	0	0	48	0	0	0	8	0	0	0	12	0	276	0	0	65	0	2	27	0	15	0	5	0	8	0	0	301	1	0	0	0	0	53	0	15	8	0	0	38	0	0	0	2	0	0	0	13
Lea Meadows	0	0	1	0	77	4	0	0	3	2	0	0	6	0	284	0	0	4	0	1	76	1	18	9	5	0	6	0	10	79	5	0	0	6	5	109	1	29	6	1	2	26	0	1	0	0	3	0	0	16
Loughborough Big Meadow	1	0	1	3	57	10	0	0	0	0	1	0	0	1	193	0	0	3	3	2	391	0	11	3	9	0	1	0	2	135	2	0	0	1	9	87	0	0	0	0	2	3	0	0	0	0	33	1	0	3
Lucas' Marsh	0	1	5	0	82	1	0	0	0	1	41	0	3	0	155	0	0	92	0	3	6	0	21	0	7	0	2	0	0	13	0	0	0	0	2	13	0	4	0	2	1	1	0	0	0	0	3	1	0	15
Lyddington Meadow	0	0	0	0	0	0	0	0	0	0	0	0	0	0	51	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Merry's Meadows	1	1	4	3	17	0	0	0	0	0	1	0	0	1	173	0	0	3	0	2	40	0	17	0	0	0	3	1	0	12	1	0	0	0	3	12	0	0	0	0	0	0	0	0	0	0	3	0	0	2
Mountsorrel and Rothley Marshes	0	0	1	0	61	0	0	0	0	0	0	0	0	0	110	0	0	1	0	1	80	0	8	0	9	0	0	0	0	92	1	0	0	0	0	7	0	0	1	0	1	0	0	0	0	0	0	0	0	2
Narborough Bog	2	0	2	2	113	8	0	0	0	2	81	0	10	0	330	0	0	123	1	5	59	1	22	1	6	0	9	0	0	310	0	1	0	0	4	131	0	30	6	0	5	31	0	0	0	0	22	1	0	15
Prior's Coppice	3	0	4	0	76	0	0	0	4	2	0	1	14	0	245	0	1	97	1	3	79	0	20	0	12	0	25	0	0	232	3	0	0	0	0	102	0	0	8	0	0	45	0	0	0	0	0	1	0	22
Rocky Plantation	0	0	0	0	15	0	0	0	0	1	0	0	0	0	24	0	0	0	0	0	2	0	1	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	11	2	0	12	0	0	0	0	0	0	0	1
Rutland Water Nature Reserve	0	1	4	0	221	0	0	0	2	6	0	1	9	0	416	0	1	107	1	3	134	0	27	0	18	0	22	1	1	393	6	0	0	0	2	129	0	56	6	1	2	28	0	1	0	0	16	0	1	28
Stonesby Quarry	2	0	1	0	24	0	0	0	0	2	0	0	1	0	189	0	0	0	0	0	100	0	19	0	1	0	6	0	0	146	1	1	0	0	1	43	0	0	0	0	4	0	0	0	0	1	0	0	3	
Syston Lake	0	0	0	0	57	2	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	12	0	2	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3	
The Miles' Piece	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Tilton Railway Cutting	0	0	0	0	8	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	11	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	
Tom Long's Nature Reserve	0	0	0	0	0	0	0	0	0	0	0	0	1	0	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Ulverscroft	2	0	4	1	100	0	1	0	6	3	0	1	6	0	205	1	1	181	7	2	167	3	23	0	2	0	16	0	5	280	2	0	1	5	5	73	1	35	12	1	5	65	0	2	0	4	51	0	0	20
Ulverscroft (Herbert's Meadow)	0	1	2	1	13	1	0	0	0	0	1	0	0	4	1	222	0	0	37	0	3	83	0	19	1	2	0	7	0	5	48	3	0	0	3	2	56	1	8	4	1	0	16	0	2	0	1	55	0	0	16
Wanlip	0	0	2	0	128	3	0	0	0	0	0	0	0	0	123	0	0	0	0	0	2	0	0	5	0	2	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	
Wanlip Rough	0	0	0	0	7	0	0	0	0	0	0	0	0	0	60	0	0	0	0	1	0	0	2	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	
Wymeswold Meadows	0	0	1	0	15	0	0	0	0	0	1	0	0	0	62	0	0	0	0	0	24	0	12	1	2	0	1	0	0	68	0	0	0	0	0	5	0	1	2	0	0	9	0	0	0	0	0	0	2	
Wymondham Rough	0	0	2	2	38	0	0	0	0	2	5	0	1	4	0	211	0	0	3	0	1	66	0	11	0	3	0	0	0	0	188	2	0	0	0	0	10	0	0	0	5	3	1	0	2	0	0	0	5	

D.5 LRWT Data Challenges

LRWT Analytic Challenges

Paul J Palmer

2019-03-21

1 Purpose

- Discuss LRWT Analytic challenges.
- Find suitable case studies for further exploration.

2 Present

- Flora & Fauna Expert
- Bats Expert
- PhD Student

3 Discussion

- Examine holdings of data and their associated analytic challenges.
- Two main type discussed:
- Analysing bat sonograms and the techniques currently used to generate reports and,
- Flora data.

3.1 Bats

Around 11,000 individually identified sonograms saved . These are manually converted into reports for public and controlled circulation using a mixture of R, Excel. QGIS and Word. This is currently a very time consuming process. There is an awareness that there is more that could be done with the data.

Analysis! Excel! Muitiple spreadsheets

Analysis! Tools, Tools! R, Tools! Excel, Tools! QGIS, Tools! Word

The analysis follows a pattern that could be incorporated in a repeatable methodology so is suitable for

3.2 Flora & Fauna

Multiple surveys available as Excel spreadsheets. Raw data in long format converted into summaries that are incorporated into reports using Word and Excel. Once again an awareness that more could be done with the data.

Analysis!Data! Long format, Data! Report! Word, Data! Summarise

Examination of the files suggests that more information could be extracted with improved techniques.

Analysis! Improve techniques, Data! Extract more information

4 Questions

End user programming experience? Very limited previous experience upon which to base adoption of R. Successful in applying cut-and-paste approach. Generally forced to use time consuming processes that will deliver a result rather than invest time in learning to program. A strong preference for graphic interface.

Analysis! Tools! R, Analysis! Programming! Limited experience

Analysis! Programming! Cut and paste, Analysis! Deliver results! No learn to program

Level of Excel skills Very high. Sample spreadsheets use filters and pivot tables to achieve tabular analysis.

Analysis! Tools! GUI preferred

Analysis! Tools! Excel! Power user

5 Follow on work

– Examine data to see if it can be incorporated into cases studies.

Analysis! Case Studies

D.6 Rutland Water NR Data WeBS

Wetland Bird Survey Data (WeBS)

Paul J Palmer

24th May 2019

1 Present

- Leicester University Fellow
- Rutland Water Reserve Manager
- PhD Student

2 Discussion

2.1 Purpose

- Explore potential link with Leicester University NERC project: NE/S009310/1. This project seeks to improve access to habitat information in and around wetland reserves.
- Discuss problems with analysing WeBS data.

Analysis! NERC, Data! Satellite imagery

Analysis! Challenge! WeBS

2.2 Background

- NERC project links satellite data to habitat information using QGIS.
- QGIS and satellite data part of project well advanced.
- Current NERC project due to end July 2019, but 8 week extension expected.
- WeBS data messy and cannot be imported into QGIS in current format.
- Can template approach help transform data on an ongoing basis?

Analysis!Data! NERC! Habitat, Data! Analysis! QGIS

Analysis!Data! WeBS! Messy, Analysis!WeBS! QGIS! Not compatible

Analysis!WeBs! Template! QGIS

2.3 Proposed actions

Create template A trial template demonstrated the feasibility of transforming the data and accommodated most most of the idiosyncrasies in the data.

Analysis! Template! WeBS, Analysis!QGIS! Template

Output format for QGIS Geospatial data-frames broken down by species agreed as the best format for the QGIS software as developed. Current R practice would prefer a single standard data-frame with geospatial information encoded as WGS84 mercator projection. Providing both formats does not represent significant extra work once the data has been transformed.

Analysis! WeBS! Geospatial

Summary output Summary outputs as chloropleth maps and five year peak averages required in the template for reporting purposes.

Analysis! Summary! Reporting, Analysis! Report! Peak averages

Missing data The trial established that only three years worth of data has been electronically provided. The historical data is required for statutory reporting on designated areas so is a high level management issue.

Analysis! WeBS! Historical data! Gaps, Analysis!WeBS! Mutiple files and formats

Mutual support –

- P J Palmer and Lboro project will be cited in NERC project as providing support to LRWT ensure that support for data transformation and import into QGIS is available beyond the NERC project. The goal being to teach others how to use the template. In this context PJP will be cited as a technically experienced volunteer deeply embedded within the organisation and management of biodiversity data in Leicestershire.

Analysis! WeBS! Template, Template! End users

- PJP will use this as a case study to demonstrate how the template approach can be used to empower organisations to take control of their own data when providing complex reports for management and statuary purposes.

Analysis!Template! Case study, Analysis! Empower end users

3 Questions

- Where is the missing data? It is possible that it is electronically encoded within the LRWT recorder system.

Analysis!WeBS! Missing data

- Paper records of the missing data have been found in a file at the LRWT head office.

Analysis!WeBS! Paper records

- Summary data has been provided by WeBS.

4 Follow on work

- Can the missing data be fully reconstituted from the electronic records?

Analysis! WeBS! Reconstitute data

D.7 LRWT Survey Review

Review survey data

Paul J Palmer

3rd May 2019

1 Present

- Survey Expert
- PhD Student

2 Discussion

2.1 Purpose

- Discuss interpretation of survey data previously provided.
- Have demonstrated template to read data, but what now?

2.2 Comments

- Very keen on R, but finding it difficult to make progress. *Analysis! Tools! R, R! Difficult*
- Hoping to be able to look at examples and then adapt them for own use. *Analysis! R! Look and learn*

3 Follow on work

- Survey data for Aylestone meadows a single survey over two years, not two surveys as originally thought.
- Other data is multiple surveys. *Analysis!Data! Muitiple surveys*
- Looking for better ways to visualise data as many landowners prefer a visual type output, rather than a report. *Analysis! Visualisations*
- Many survey undertaken for private land owners seeking to improve their wildlife land management practice. *Data! Private, Analysis! Management practice*

- Species diversity index, while reducing survey to a single number, does reflect expert opinion.

Analysis! Species Diversity Index

D.8 Rutland Water NR Species Inventory

Rutland Water Species Inventory

Paul J Palmer

4th July 2019

1 Discussion

1.1 Purpose

- Invited presentation on the species inventory based on the LRWT data holdings.
- Part of the RW volunteers recording meetings
- Two presentations:
  1. Sue Timms on electronic recording flow
  2. PJP Species inventory
- Well matched pair of presentations with one looking and the flow and storage and the other looking at data.

1.2 Present

Invited Recorders RW and other natural history societies.

PhD Student As RW recorder and PhD researcher.

1.3 Background

Presentation based on an analysis of LRWT recorder data clipped to include only the records within a bounding box of the Rutland Water area. This process will only include records for which a correct geospatial location is included.

The data had gaps, especially for birds. This was mentioned independently by ST in her presentation.

Everyone believes that the data exists, probably as Excel spreadsheets, or maybe in Map-Mate, and unsupported Windows based program.

2 Questions

- What are the gaps in records? *Data! WeBS! Gaps*
- Where are the missing records? *Data! Data holders*
- Are the bird records held by LOROS, and if so, will they make them available? *Data! WeBS! Access*

3 Follow on work

- Add data from ORCA *Data! ORCA*
- Accept ST offer to upgrade access to ORCA
- Add additional data from NatureSpot as ORCA might not include latest sightings.
- A holy grail for county recorders is the ability to rapidly produce an atlas of species records. *Analysis! Aspiration! Atlas*

D.9 LRERC Stakeholder Interview

LRERC

Paul J Palmer

31st July 2019

1 Present

- LRERC Ecologist
- LRERC Ecologist
- PhD Student

2 Discussion

2.1 Purpose

- Discuss how reusable templates might be used to help with real LRERC problems.
- Learn more about the data issues at LRERC.

2.2 Background

- Formatting place names into a consistent hierarchy is necessary since long text might be trimmed. Place names should be presented in coverage order. County; Town; Locality. Since some software limits the length of text data then ends can get trimmed. Using this method ensures that the first part of the name is preserved. *Data! placenames*
- Many records have a post code associated which should not be present and might be misused to identify an individual. *Data! Privacy*
- GDPR principles are regarded as sensible by professional data managers to prevent inappropriate identification of individuals. *Data! Privacy! GDPR*
- Manual perusal of records is the only way that personal information is cleaned. *Data! Privacy! Manual Checking*

- Many local groups keep records in formats such as map-mate excel and even Access. There are inconsistencies between even in the way personal names are entered. *Data!Synonyms, Tools! Map-Mate, Tools! Excel, Tools! Access*
- Errors in data are common including incorrect place names. *Data! Errors*

3 Questions

- Management of personal information within data:
  - Can address components be identified and automatically removed? *Aspiration! Identify errors*
  - Can structured location information be applied using mapping information? *Aspiration! Add Location*
- Can LRERC install and use R so that they might use templates *Aspiration! Use R*
- Can atlas making be generalised to motivate local recording groups? *Aspiration! Atlas making*

4 Follow on work

- Try identifying personal data that should be removed in AR moth data. *Aspiration! Remove personal data*
- Try producing atlas type layouts with map backgrounds. *Aspiration! Atlas making*

D.10 End User R Analysis

Notes from enduser contribution to R analysis

Paul J Palmer

15th September 2020

1 Present

- NatureSpot Trustee
- PhD Student

2 Discussion

2.1 Purpose

- Creation of reports form NatureSpot data. *Analysis! Nat*

2.2 Background

These charts were volunteered by Alan:

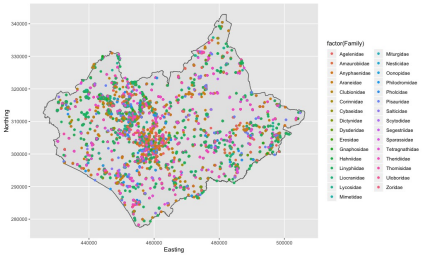


Figure 1: A first R plot

Although a good start, these figures had technical problems with their use:

- Spaces in names, so not acceptable to Latex. *Analysis! R!, Issues! syntax*
- In JPEG format, so low quality
- Excessive white space around graphic.

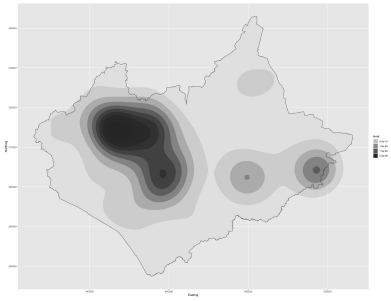


Figure 2: 43k records heatmap

3 Questions

- Encourage use of R markdown to get reports?

4 Follow on work

- an item

D.11 LRWT Bat Analysis Requirements

Notes from Email communication

Paul J Palmer

19th August 2019

1 Actors

- LRWT Conservation Officer
- PhD Student

The following text has been collated from social media by NC and relates to the challenges of analysing data collected from bat surveys. Theses surveys use recording equipment to transform the ultrasonic calls from bats into the human audible range where they may be identified. Typically the annotation process in done by hand resulting large data that requires some post processing before analysis.

By way of background PJP has already made a 'proof of principle' template for bat data analysis which now needs further development and transforming into more elegant format so that Nathalie may trial it.

2 Email extract

Personal names commented out. Minor edits made for readability.

One for the hive mind: are there any scripts for R that people could recommend for statistical analysis of bat survey data for both static detector deployments and transects. Thank you for any help in advance.

Don't have any links. When you say statistical analysis do you mean actual statistical analysis or just descriptive stats? Most bat surveys are not designed thorough enough for meaningful analysis.

Everett At this stage descriptive stats is what I want to look at. I am just starting out with R.

A quick google search brought up this why will probably help you out:
Link to Bat Survey Guidelines. This document includes advice on how to incorporate statistically valid elements into a survey of this type.

I had found that one previously and was wondering if there were any scripts people were using?

What do you actually want? `Ecobat.org.uk` will analyse data for you (running R in background). Ecobat provides tools for the standardised, rigorous interpretation of bat activity data.

I want to get to grips with R and as I collect bat data I was looking to use this to help with learning R as I think it's easier to learn something with actual data. I have seen the Ecobat site and will be using it.

I have plenty of scripts but as you said, it really depends on what you want to do. Actograms, activity per day compare sites, of make maps? I can send you some scripts but those are tailored to batcorder exports...

I'd be interested in having some of those scripts, if that's okay. I'll probably end up writing my own but I like to have a working base to work from. I'll be comparing sites and making maps mainly. Don't worry about the fact that's it's tailored for batcorders, I can most likely deal with that.

I would be interested in the scripts to have a look at.

I am happy to send you scripts but do not want to swamp you, and my scripts may not make much sense when you try to run them on your data. R is so versatile that you really need to decide what you want to test/visualise. A nice start might be:
Weissgerber, T. L. et al. (2015) 'Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm', PLoS Biology. Public Library of Science, 13(4), p. e1002128. doi: 10.1371/journal.pbio.1002128.

I am actually interested in how people code. Everyone does it differently and I like to learn from that. I understand the theory and know my way around the basics of R. What I'm really interested in here is how people write their code... I code stealthily. :-)

3 Discussion

- This social media thread provides empirical evidence in a general interest in R that is stimulated by the amount of data generated by current bat survey techniques.

4 Questions

- The charity EcoBat provides an online tool for the standardised analysis of bat data. Does this tool meet the needs of the community?
- Are data stakeholder interested in online tools, or do they want to own the whole analysis?

D.12 Access To Sports Science Data

Sports Science Data

Paul J Palmer

21st August 2019

1 Present

- Supervisor.
- School of Sport, Exercise and Health Sciences.
- PhD Student.

2 Discussion

2.1 Purpose

- Discuss potential for using reusable templates to help collate sports science data.

2.2 Discussion

- Sports science data is produced all around the world so there is a need to measure and combine data from many sources.
- The need is described as: 'harmonise' data, which requires an understanding of the 'correlates' within the data to ensure an appropriate transformation.
- Correlates include:
  - Age
  - Gender
  - Height
  - Car usage
  - Other social factors
  - health factors
- Problems with harmonisation include:
  - Incorporation of retrospective data
  - Units
  - Temporal factors e.g. (Multiple wave cohorts)
  - Categorisations
  - Context specific interpretation of categorisations
- Tools used
- Data format and repository

- Repository for harmonised data ICAD.
  http://www.mrc-epid.cam.ac.uk/research/studies/icad/
- Raw accelerometer data is released to ICAD. Study specific data goes through the harmonisation process.
- Raw accelerometer data through a specific software 'KineSoft' and produces outputs as a batch of Excel spreadsheets.
- Estimate about 60,000 participants and 80 datasets of which 40 have been harmonised. There are about 2000 variables in a dataset.
- Excel a key tool for external manipulation of data.
- Original data and STATA always preserved for confirming provenance of harmonised data.
- STATA
  https://www.stata-uk.com/software/stata.html/?utm_medium=adwords&utm_campaign=statauk&utm_source=software&gclid=EAIaIQobChMIj8zt59il5AIVTbTtCh0RyAfqEAAYASAAEgJ-E_D_BwE

Two methods papers for ICAD which may help to clarify some of the methods described:
Sherar, L. B. et al. (2011) 'International children's accelerometry database (ICAD): Design and methods', BMC Public Health, 11(1), p. 485. doi: 10.1186/1471-2458-11-485.
https://www.ncbi.nlm.nih.gov/pubmed/21693008
Atkin, A. J. et al. (2017) 'Harmonising data on the correlates of physical activity and sedentary behaviour in young people: Methods and lessons learnt from the international Children's Accelerometry database (ICAD)', International Journal of Behavioural Nutrition and Physical Activity. BioMed Central, 14(1), p. 174. doi: 10.1186/s12966-017-0631-7.
https://www.ncbi.nlm.nih.gov/pubmed/29262830

The second one (Atkin et al.) is more pertinent to the discussions. Additional information about the data harmonisation can be found here:
http://www.mrc-epid.cam.ac.uk/research/studies/icad/data-harmonisation/

3 Questions

- Are R and RStudio an acceptable tool to adopt? Yes!

4 Follow on work

- Exploratory meeting with other stakeholders.

D.13 Observations From R Course

Sports Science Data

Paul J Palmer

13th September 2019

1 Present

25 Course attendees

PhD Student

2 Discussion

2.1 Purpose

- R training course focussed on working with multiple data sources.
- Social science data relation to life expectancy in London boroughs used in group exercise.
- Main challenge presented in the course was the joining of multiple data to make a single dataset

2.2 Background

Course organised by National Centre for Research Methods (NCRM) under the banner of advancing social science research methods and held at Manchester University.

3 Comments from attendees

- Attendees commented that they and other R users often had to work on personal computers as organisational computer services did not support R.
- One user commented that although R was supported, self download of packages was not, so users had to agree in advance which packages were pre-downloaded.
- Another user commented that for courses he set up local libraries which circumvented these restrictions. (Does this work generally?)
- See note below about partial installation of R Studio.
- It seemed that R Markdown and R packages were not used by anyone else because they are seen as intimidating.
- A user asked if I had systems training because of my approach to using R. The same user attended multiple R courses and said that much could be learned from other attendees, rather than from the course.

4 Observations about course

- Course notes were circulated on a memory stick rather than as a package. Preparation of student course material is cited as one of the reasons for using packages on user help fora.
- The R Studio installation was incomplete. Only R & R Studio were installed. Latex and Pandoc were missing, preventing the use of R markdown documents so literate programming principles could not be used. This was surprising since the source book for the course notes, 'R for Data Science', recommends the use of R Markdown as part of the communication process.
- Overall R was treated like a programming language rather than a language of analysis.
- Loose definitions used for long and wide data that did not quite match those normally used in books on TidyR.
- Choice of some sample code was odd; for example, using lengthy commands to rename columns rather than simple base R commands. When queried the answer was: 'We are not programmers.'
- No account was made for non UTF8 character sets and the built-in tools for addressing charset issues.
- No explanation was made regarding the (sensible) use of a list of lists to share sample data, rather than the data frames and tibbles described in the lecture notes. This probably led to some of the difficulty many delegates had with accessing data.
- Overall I appeared to be the most advanced R user present, especially with regards to the preparation of data. Conversely many delegate were much more familiar with linear models and their application than me.

D.14 Comments On R From Sheffield University

Email: R Usage At Sheffield University

Paul J Palmer

15th September 2020

1 Addressees

- To PhD Student
- From: Senior Lecturer, Sheffield University.

2 Email Body

Slightly edited for jargon and redaction of non-relevant personal material.

2.1 Enquiry

I am currently working on my PhD at Loughborough University and looking at ways of building templates in R to facilitate the analysis of large data. I have been particularly focussed on the challenges of initially reading and cleaning the data, since this is often the first barrier met by less programatically adept users. Will mentioned that you have a well established culture of supporting R at Sheffield University, which is in contrast to the stories I have collected where users are meeting resistance to its use.

I would find it very helpful to be able to quote examples of organisations that support the use of R, so wonder if you could help?

2.2 Reply

The history of using R in the SMI (Sheffield Methods Institute) is largely that we've not existed for that long - unlike most university departments or institutes (as I'm sure you well know) with long histories and accompanying existing practices, we've only existed since 2014, when a bunch of new people were hired. As a consequence of that we've been able to build something from the ground up, and those of us who teach quants (Quantitative Methods) all more-or-less agreed that we'd prefer to teach with R than anything else

While there's a lot of resistance to taking on R, there's also a bunch of organisations that have been supportive of R - the main setting in which this is collected is actually RStudio, who - given their commercial arm - are a bit more proactive at talking about it than, say, the R Consortium. There's some case studies on their own website [1], There's also some examples that I've heard about anecdotally, like StitchFix [2] (especially Hilary Parker's [3] work) and Uber. [4]

But there's a lot out there, especially using tidyverse [5] tools (given, as you mention, the amount of time spent on data cleaning and manipulation).

[1] https://www.rstudio.com/about/customer-spotlight/
[2] https://multithreaded.stitchfix.com/algorithms/
[3] https://hilaryparker.com/
[4] https://eng.uber.com/dsw/
[5] https://www.tidyverse.org/

3 Supplementary Questions

3.1 Is R used as a primary tool?

I followed the link on your home page to the Q-Step programme. If I understand what I read correctly, this is promoting quantitive numeracy in the social science domain. The documents make a passing, but positive reference to R and seem to suggest that only a few centres are teaching R as a primary tool. From what you have told me one such centre is you, but are there any other Q-Step centres taking the same stand on R as you?

The second question is straightforward: Have you authored any papers where R has been the principle tool used for the analysis?

Sheffield University don't seem to make it easy for some one outside of the University to find publications, but "Nonparticipation or different styles of participation? Alternative interpretations from Taking Part" seemed to suggest that STATA was used as the primary tool

3.2 Reply

I'm not sure about all the Q-Step centres, to be honest. We're definitely not the only one - I know Glasgow is taking a similarly firm line to us - and there's others that are using at least some R in their teaching (Manchester's a good example), but I can't generalise too strongly.

Because as you say Sheffield's not very good at making it easy to find papers, it'll be quicker to go via my Google Scholar page than my university page. If you filter by stuff in the last couple of years, you'll see that a lot of it is a mix of Stata and R depending on the task, but this one's [6] all R with nothing else.

3.3 What type of computers used?

A final question after reading your paper 'The coming crisis of cultural engagement? Measurement, methods, and the nuances of niche activities', which I found an engaging read and a nice change from my usual diet of engineering papers. You analysed some quite large data (in excess of 33 million rows) - did you use an ordinary PC or Mac for this task?

3.4 Reply

Both - I use a Windows machine in the office and have an Apple laptop, and the code ran fine on both, whether I was working on which one was largely a function of where I was at the time. (The main thing was the data.table package, which makes reading in enormous files much more straightforward - once the data was in the analysis wasn't too bad. My general measure is "is my computer getting appreciably hot", which it wasn't)

4 Comments

- With the exception of Uber, I had previously encountered all of the other URLs.
- There is a particular default style to ggplot generated graphs so I had guessed that a mixture of softwares were used including R. However, themes can easily be overridden and any arbitrary theme emulated.
- 33 million rows of data fits in with my view that 100 rows plus should be possible on a pc.

[6] Hanquinet, L., O'Brien, D. and Taylor, M. (2019) 'The coming crisis of cultural engagement? Measurement, methods, and the nuances of niche activities', Cultural Trends. Routledge, 28(2–3), pp. 198–219. doi: 10.1080/09548963.2019.1617941.

D.15 NatureSpot Analysis

Email:

Paul J Palmer

20th October 2019

1 Addressees

- To PhD Student
- From NatureSpot Trustee.

2 Email

2.1 Action

A brief discussion at the Leicestershire Entomological Society meeting on the 18th October 2019 about what observations might be drawn annually from NatureSpot data prompted the sharing of the following email and Excel spreadsheet.

2.2 Email Body

Slightly edited for jargon and redaction of non-relevant personal material.

I've attached a copy of my spreadsheet with all the VC55 mollusc data. There are a lot of tabs which can probably ignore. The first tab contains the raw data and the last two are where I did the calculations. You can no doubt achieve this is a much simpler way with code!

On the last tab you will see the status column which I filled in beforehand as a comparator. Of course the four category boundaries can be set wherever you want but I thought there was a pretty good match basing the status on the record .

There are some curly bits when it comes to thinking about rarity, as some species have more restricted habitat requirement so are naturally less widespread. I haven't yet looked at the trends (ie percentage of records within the taxon per year) as I suspect the dataset is too small for most species.

I have been thinking about trying this approach with the bird data to see how that looks (I can pull the bird dataset off if you are interested). I would really like to be able to show that we are drawing some conclusions from the (NatureSpot) data!

Let me know how you get on.

D.16 NatureSpot Paper Planning

NatureSpot Paper Planning

Paul J Palmer

16th December 2019

1 Present

- Paper authors.

2 Discussion

2.1 Purpose

- Planning content for paper targeted in response to the BES Journals 'Citizen Science special issue call' for People and Nature.
  `https://besjournals.onlinelibrary.wiley.com/hub/journal/13652656/special-features`.
- Draft paper is available on Overleaf. (Editing rights required for access):
  `https://www.overleaf.com/project/5dcd210c35e22800019c5ef1`
- Submission date: 21st January 2020.

2.2 Background

- The motivation for this paper is build a better understanding of the biodiversity records collated by NatureSpot with the end goal of better reporting of those records along with a better understanding of VC55 biodiversity.
- Specifically this paper is based on an analysis of NatureSpot data with an emphasis on understanding the behaviour of contributors, rather than the reported biodiversity records.
- Understanding the contributor behaviour should allow a better understanding of recorder bias and support a more refined understanding of the underlying biodiversity.
- It is a duty for public authorities to conserve biodiversity and schemes such as NatureSpot may offer supporting evidence:
  `https://www.gov.uk/guidance/biodiversity-duty-public-authority-duty-to-have-regard-to-conserving-biodiversity`

3 Questions

- Why are there white holes in the data?
- Tetrad bashing reveals that some of the white holes are not too bad for biodiversity.
- Can we see a correlation between records and:
  - Post Codes (proxy for population).
  - Roads (proxy for access).

Appendix E

Interview Qualitative Data Analysis

QDA Analysis

Paul J Palmer

2020-03-25

1 Qualitative Data Analysis Analysis Of End User Interview Notes

This analysis uses three related sources of data: meeting notes annotated with QDA codes; highlighted sections covered by those codes; and complete pdf files of the meeting notes. The annotations and highlighted text are subjectively chosen by the author as significant and marked with author selected keywords. Each source of data are collated from a sub-directory containing the meeting notes into a single corpus which allows an *ad hoc* approach to updating the source data without the need to rewrite the analytical code. This approach has been inspired by the reproducible and reusable themes which are the subject of this research. It is also necessary to remove common words that contribute little to the understanding extracted by this analysis. A list of common English "stop words" are provided by the *tidytext* R package and subtracted from the corpus. These have been enhanced by subtracting words from the *pro forma* document template used to record interviews along with manually selected words such as personal pronouns.

1.1 Method

The meeting notes have been marked up using the Latex *ulqda* package which allows the insertion of custom codes (singly or hierarchically) into the document with highlighting to indicate the portion of text to which it applies. The *ulqda* package was chosen as all meeting note were made using Latex and the package produces CSV file which are convenient for analysis in R and Markdown. This document has been generated by following the literate programming support provided by the R KnitR package and therefore is a simple example of reproducible analysis. It is reusable in the simplistic sense that CSV data is read from directories adjacent to the directory holding this markdown file using a recursive search process. Thus, additional adjacent directories may be added, the notes within coded, and then included in this analysis by knitting the document again.

Subsequent analysis can focus on either the coding or the highlighted text, minus the stop words. The two sources allow for a certain amount of cross checking of results, although there is no independence since both notes and coding where performed by the same researcher. It is also possible to perform an un-coded analysis directly on PDF documents by extracting the words from the document. In this case, the subtraction of words used in pro forma styles is particularly important as these will appear in all documents. Care has been taken to ensure that only the target files have been selected for inclusion through appropriate naming of files and regex filters. All meeting notes have a filename that begins YYYY-MM-DD and the pro-forma documents include this term in the name. Since the data are under control of the researcher this make for a convenient simplification in program code, but it does require manual checking to ensure that files are not accidentally included or omitted.

1.2 Analysis of QDA Code Annotations

The word cloud Figure 1 is derived from the codes used to annotate the meeting notes. The words highlighted: data; analysis; tools; management; aspiration, agree with a sense of those areas which are important to the stakeholders that were picked up during informal discussions. The desire to analyse data for management

Figure 1: Wordcloud from code annotations

purposes is clear, but words in the background also resonate: slow; manual; duplicates; errors; messy; cleaning, are all indicative of perceived problems. Whereas; reporting; monitor; targets; digitisation; atlas, all suggest aspirational goals that can be addressed through data analysis.

Performing a similar analysis on those areas of text highlighted during the coding process will give a cross check on the validity of the coding summary.

1.3 Analyse highlighted text

Figure 2: Wordcloud from highlighted text

The wordcloud formed from the highlighted text Figure 2 has an similar emphasis on management and data issues. Also notable are references to tools routinely used: orca (county level database); excel (spreadsheet); qgis (open source geospatial software); database; portal. There is no doubt from the context of discussions that the focus of these tools are on the the summary analysis of data for reporting purposes, and words such as 'manually' and 'raw' which relate to the difficulty of achieving this analysis. Also apparent are references

to recording process with terms such as: recorder; volunteer; and survey, which were not used in the QDA terminology.

2 Using PDF As A Corpus.

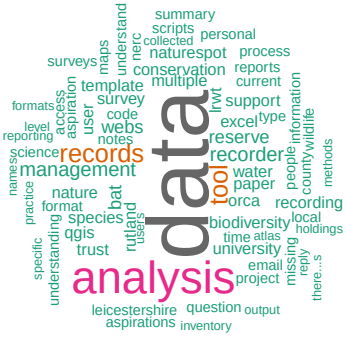


Figure 3: Wordcloud from pdf notes

Using the collated PDFs as a corpus has the advantage of convenience, in that it would also work with un-coded documents. The wordcloud in Figure 3 shows strong similarities to Figure 2 confirming that no additional terminology needs to be considered for inclusion.

2.1 Network view of bigrams

Since word clouds highlight frequency of word use, they do not give any sense of relationship between the terms, but a network based on adjacency of words helps to address this issue.

The network view in Figure 4 helps to show the relationships between word groups by listing cases where adjacent words occur at least twice in the text corpus. Some knowledge of the domain is needed to interpret the threads. For example, an aspirational goal is the production of a species atlas and inventory. There are a number of different analytical challenges. Given that the core theme is data monitoring and management is important for the conservation of nature reserves.

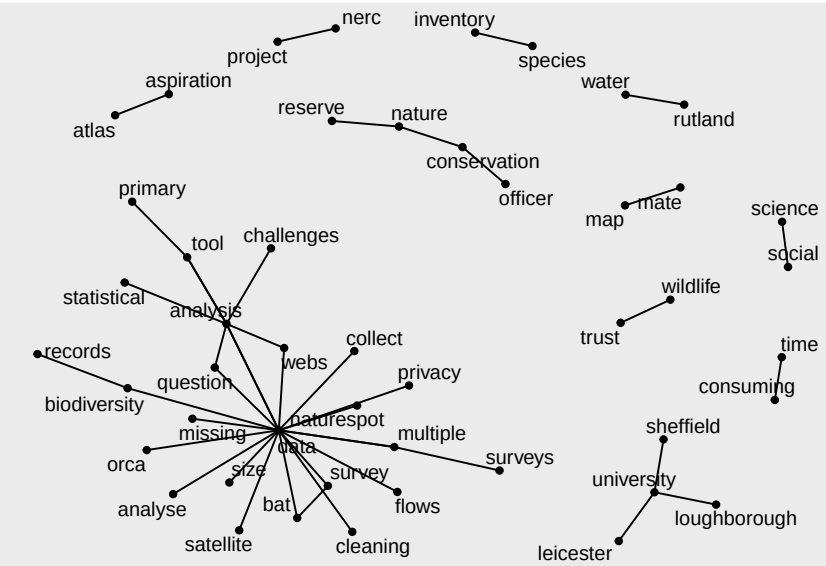


Figure 4: Network of keywords from pdf data.

Appendix F

Draft Publications

No papers were accepted for publication during this research, but three are in draft and two are available on non peer reviewed preprint servers. Self publication on these servers will not affect future acceptability for journals.

F.1 A Modular Task Orientated Approach For The Analysis Of Large Datasets

https://doi.org/10.31219/osf.io/ys2vw

F.2 Does Citizen Science Biological Recording Tell Us As Much About The Recorders As Biodiversity?

https://doi.org/10.31219/osf.io/bsye5

F.3 Beyond Maps: Visualising Citizen Science Biodiversity Data With Open Source Tools

Submitted to Journal of Applied Ecology and Evolution.