Contents

List of Figures xiii

Preface xv

Author xvii

CHAPTER 1 ▪ Introduction 1

1.1 DATA STEWARDSHIP FOR OPEN SCIENCE 1

1.2 INTRODUCTION BY THE AUTHOR 6

1.3 DEFINITIONS AND CONTEXT 29

1.4 THE LINES OF THINKING 36

1.5 THE BASICS OF GOOD DATA STEWARDSHIP 38

CHAPTER 2 ▪ Data Cycle Step 1: Design of Experiment 63

2.1 IS THERE PRE-EXISTING DATA? 63

2.2 WILL YOU USE PRE-EXISTING DATA (INCLUDING OPEDAS)? 65

2.3 WILL YOU USE REFERENCE DATA? 66

2.4 WHERE IS IT AVAILABLE? 68

2.5 WHAT FORMAT? 69

2.6 IS THE DATA RESOURCE VERSIONED? 70

2.7 WILL YOU BE USING ANY EXISTING (NON-REFERENCE) DATASETS? 72

2.8 WILL OWNERS OF THAT DATA WORK WITH YOU ON THIS STUDY? 73

2.9 IS RECONSENT NEEDED? 74

3.8 COMPUTE CAPACITY PLANNING 143
3.8.1 Determine needs in memory/CPU/IO ratios 143

CHAPTER 4 ▪ Data Cycle Step 3: Data Capture (Equipment Phase) 147

4.1 WHERE DOES THE DATA COME FROM? WHO WILL NEED THE DATA? 150
4.2 CAPACITY AND HARMONISATION PLANNING 152
4.2.1 Will you use non-equipment data capture (i.e., questionnaires, free text)? 153
4.2.1.1 Case report forms? 155

CHAPTER 5 ▪ Data Cycle Step 4: Data Processing and Curation 157

5.1 WORKFLOW DEVELOPMENT 157
5.1.1 Will you be running a bulk/routine workflow? 158
5.2 CHOOSE THE WORKFLOW ENGINE 159
5.2.1 Who are the customers that use your workflows? 162
5.2.2 Can workflows be run remotely? 163
5.2.3 Can workflow decay be managed? 164
5.2.4 Verify workflows repeatedly on the same data 165
5.3 WORKFLOW RUNNING 166
5.4 TOOLS AND DATA DIRECTORY (FOR THE EXPERIMENT) 168

CHAPTER 6 ▪ Data Cycle Step 5: Data Linking and 'Integration' 171

6.1 WHAT APPROACH WILL YOU USE FOR DATA INTEGRATION? 175
6.2 WILL YOU MAKE YOUR OUTPUT SEMANTICALLY INTEROPERABLE? 176
6.3 WILL YOU USE A WORKFLOW OR TOOLS FOR DATABASE ACCESS OR CONVERSION? 179

CHAPTER 7 ▪ Data Cycle Step 6: Data Analysis, Interpretation 181

7.1 WILL YOU USE STATIC OR DYNAMIC (SYSTEMS) MODELS? 182

7.2 MACHINE-LEARNING? 183
7.3 WILL YOU BE BUILDING KINETIC MODELS? 185
7.4 HOW WILL YOU MAKE SURE THE ANALYSIS IS BEST SUITED TO ANSWER YOUR BIOLOGICAL QUESTION? 186
7.5 HOW WILL YOU ENSURE REPRODUCIBILITY? 187
7.5.1 Will this step need significant storage and compute capacity? 188
7.5.2 How are you going to interpret your data? 189
7.5.3 How will you document your interpretation steps? 191
7.6 WILL YOU BE DOING (AUTOMATED) KNOWLEDGE-DISCOVERY? 192

CHAPTER 8 ▪ Data Cycle Step 7: Information and Insight Publishing 195

8.1 HOW MUCH WILL BE OPEN DATA/ACCESS? 196
8.2 WHO WILL PAY FOR OPEN ACCESS DATA PUBLISHING? 199
8.3 LEGAL ISSUES 202
8.3.1 Where to publish? 203
8.4 WHAT TECHNICAL ISSUES ARE ASSOCIATED WITH HPR? 206
8.4.1 What service will be offered around your data? 207
8.4.2 Submit to an existing database? 209
8.4.3 Will you run your own access Web service for data? 210
8.4.4 How and where will you be archiving/cataloguing? 212
8.5 WILL YOU STILL PUBLISH IF THE RESULTS ARE NEGATIVE? 213
8.5.1 Data as a publishable unit? 214
8.5.2 Will you publish a narrative? 215

Bibliography 219

Index 223