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

2.10 DO YOU NEED TO HARMONIZE DIFFERENT SOURCES OF OPEDAS? 75

2.11 WHAT/HOW/WHO WILL INTEGRATE EXISTING DATA? 76

2.11.1 Will you need to add data from the literature? 77

2.11.2 Will you need text-mining? 78

2.11.3 Do you need to integrate or link to a different type of data? 79

2.12 WILL REFERENCE DATA BE CREATED? 81

2.12.1 What will the IP be like? 82

2.12.2 How will you maintain it? 84

2.13 WILL YOU BE STORING PHYSICAL SAMPLES? 85

2.13.1 Where will information about samples be stored? 87

2.13.2 Will your data and samples be added to an existing collection? 88

2.14 WILL YOU BE COLLECTING EXPERIMENTAL DATA? 90

2.15 ARE THERE DATA FORMATTING CONSIDERATIONS? 91

2.15.1 What is the volume of the anticipated dataset? 92

2.15.2 What data formats do the instruments yield? 93

2.15.3 What preprocessing is needed? 94

2.15.3.1 Are there ready-to-use workflows? 95

2.15.3.2 What compute is needed? 96

2.15.4 Will you create images? 98

2.16 ARE THERE POTENTIAL ISSUES REGARDING DATA OWNERSHIP AND ACCESS CONTROL? 99

2.16.1 Who needs access? 100

2.16.2 What level of data protection is needed? 101

2.16.2.1 Is the collected data privacy sensitive? 102

2.16.2.2 Is your institutes' security sufficient for storage? 103

CHAPTER 3 ▪ Data Cycle Step 2: Data Design and Planning 105

3.1 ARE YOU USING DATA TYPES USED BY OTHERS, TOO? 106

3.1.1 What format(s) will you use for the data? 107

3.2 WILL YOU BE USING NEW TYPES OF DATA? 108

3.2.1 Are there suitable terminology systems? 109

3.2.2 Do you need to develop new terminology systems? 111

3.2.3 How will you describe your data format? 112

3.3 HOW WILL YOU BE STORING METADATA? 114

3.3.1 Did you consider how to monitor data integrity? 115

3.3.2 Will you store licenses with the data? 117

3.4 METHOD STEWARDSHIP 119

3.4.1 Is all software for steps in your workflow properly maintained? 120

3.5 STORAGE (HOW WILL YOU STORE YOUR DATA?) 123

3.5.1 Storage capacity planning. 123

3.5.1.1 Will you be archiving data for long-term preservation? 124

3.5.1.2 Can the original data be regenerated? 126

3.5.1.3 If your data changes over time, how frequently do you do backups? 127

3.5.2 When is the data archived? 129

3.5.3 Re-use considerations: Will the archive need to be online? 130

3.5.4 Will workflows need to be run locally on the stored data? 131

3.5.4.1 Is there budget to enable supported reuse by others (collaboration/co-authorship)? 133

3.5.5 How long does the data need to be kept? 135

3.5.6 Will the data be understandable after a long time? 137

3.5.7 How frequently will you archive data? 138

3.6 IS THERE (CRITICAL) SOFTWARE IN THE WORKSPACE? 139

3.7 DO YOU NEED THE STORAGE CLOSE TO COMPUTER CAPACITY? 141

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