

Guidance framework to apply ~~good practice~~ best practices in ecological data analysis: Lessons learned from building Galaxy-Ecology

Royaux Coline^{1,2*}, Mihoub Jean-Baptiste³, Jossé Marie⁴, Pelletier Dominique⁵, Norvez Olivier⁶, Reecht Yves^{7,8}, Fouilloux Anne⁹, Rasche Helena¹⁰, Hiltemann Saskia¹¹, Batut Bérénice^{12,13}, Eléaume Marc^{14,15}, Seguineau Pauline^{14,15}, Massé Guillaume¹⁶, Amossé Alan¹⁷, Bissery Claire^{8,18}, Lorrilliere Romain³, Martin Alexis¹⁹, Bas Yves^{3,20}, Virgoulay Thimothée^{21,22}, Chambon Valentin¹⁷, Arnaud Elie², Michon Elisa²³, Urfer Clara^{2,24}, Trigodet Eloïse^{21,24}, Delannoy Marie³, Loïs Gregoire³, Julliard Romain³, Grüning Björn²⁵, The Galaxy-E community, Le Bras Yvan²

¹ UMR8067 Biologie des Organismes et Ecosystèmes Aquatiques (BOREA, MNHN-CNRS-SU-IRD-UCN-UA), Sorbonne Université, Station Marine de Concarneau - Concarneau, France

² Pôle national de données de biodiversité, UAR2006 PatriNat (OFB-MNHN-CNRS-IRD), Muséum National d'Histoire Naturelle, Station Marine de Concarneau - Concarneau, France

³ Centre d'Écologie et des Sciences de la Conservation (UMR7204 CESCO, MNHN-CNRS-SU), Muséum National d'Histoire Naturelle, Sorbonne Université, Centre National de la Recherche Scientifique - Paris, France

⁴ Data Terra, Centre National de la Recherche Scientifique - Brest, France

⁵ UMR DECOD (Ifremer-Agrocampus Ouest-INRAE) - Lorient, France

⁶ Pôle National de Données de Biodiversité, UAR2006 PatriNat (OFB-MNHN-CNRS-IRD), Fondation pour la Recherche sur la Biodiversité, Muséum national d'Histoire naturelle - Paris, France

⁷ Institute of Marine Research - Bergen, Norway

⁸ Institut français de recherche pour l'exploitation de la mer (Ifremer) - Brest, France

⁹ Simula Research Laboratory - Oslo, Norway

¹⁰ Department of Pathology and Clinical Bioinformatics, Erasmus Medical Center - Rotterdam, The Netherlands

¹¹ Institute of Pharmaceutical Sciences, Faculty of Chemistry and Pharmacy, University of Freiburg - Freiburg, Germany

¹² Institut Français de Bioinformatique, CNRS UAR3601 - Évry, France

¹³ Mésocentre, Clermont-Auvergne, Université Clermont Auvergne - Clermont-Ferrand, France

44 ¹⁴ Institut de Systématique Evolution, Biodiversité (UMR7205 ISYEB, MNHN-CNRS-SU-
45 EPHE), Département Origines et Évolution, Muséum national d'Histoire naturelle - Paris,
46 France
47 ¹⁵ Institut de Systématique Evolution, Biodiversité (UMR7205 ISYEB, MNHN-CNRS-SU-
48 EPHE), Département Origines et Évolution, Station Marine de Concarneau - Concarneau,
49 France
50 ¹⁶ UMR LOCEAN (CNRS-SU-IRD-MNHN), Centre National de la Recherche Scientifique,
51 Station Marine de Concarneau - Concarneau, France
52 ¹⁷ Muséum National d'Histoire Naturelle, Station Marine de Concarneau - Concarneau,
53 France
54 ¹⁸ Université Claude Bernard Lyon 1 - Lyon, France
55 ¹⁹ UMR8067 Biologie des Organismes et Ecosystèmes Aquatiques (BOREA, MNHN-CNRS-
56 SU-IRD-UCN-UA), Muséum national d'Histoire naturelle - Paris, France
57 ²⁰ UAR2006 PatriNat (OFB-MNHN-CNRS-IRD), Muséum national d'Histoire naturelle - Paris,
58 France
59 ²¹ Centre d'Écologie et des Sciences de la Conservation (UMR7204 CESCO, MNHN-CNRS-
60 SU), Muséum National d'Histoire Naturelle - Concarneau, France
61 ²² Université de Montpellier - Montpellier, France
62 ²³ Institut des Sciences de la Mer de Rimouski, Université du Québec à Rimouski -
63 Rimouski, Québec, Canada
64 ²⁴ Université de Bretagne Occidentale - Brest, France
65 ²⁵ Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University
66 Freiburg - Freiburg, Germany

67
68 *Corresponding author
69 Correspondence: coline.royaux@mnhn.fr
70

71 **ABSTRACT**

72 Numerous conceptual frameworks exist for ~~good practice~~best practices in
73 research data and analysis (e.g. Open Science and FAIR principles). In
74 practice, there is a need for further progress to improve transparency,
75 reproducibility, and confidence in ecology. Here, we propose a practical
76 and operational framework for researchers and experts in ecology to
77 achieve ~~good practice~~best practices for building analytical procedures
78 from individual research projects to production-level analytical pipelines-
79 ~~based on atomisation and generalisation~~. We introduce the concept of
80 atomisation-to identify analytical steps which support generalisation by
81 allowing us to go beyond single analyses. The term atomisation is
82 employed to convey the idea of single analytical steps as “atoms”
83 composing an analytical procedure. When generalised, “atoms” can be
84 used in more than a single case analysis. These guidelines were
85 established during the development of the Galaxy-Ecology initiative, a
86 web platform dedicated to data analysis in ecology. Galaxy-Ecology allows
87 us to demonstrate a way to reach higher levels of reproducibility in
88 ecological sciences by increasing the accessibility and reusability of
89 analytical workflows once atomised and generalised.
90

98 Ecology's Reproducibility Crisis

99 Research in ecology is increasingly shaped by the availability of novel
100 analytical solutions and statistical tools. Given the ever-growing amount of
101 data available, much attention is often given to the thought process behind
102 statistical analyses to handle different data distributions, pseudo-replication,
103 and sampling biases for instance (NERC 2010, 2012; Hampton *et al.*, 2017;
104 Emery *et al.*, 2021). Despite the high-quality standards required by the
105 scientific community from data access to analysis, the level of complexity of
106 ecological systems makes results difficult to reproduce. The ongoing
107 "reproducibility crisis" has also led researchers to pay closer attention to the
108 quality of analyses to increase confidence in their studies and conclusions
109 (Ioannidis, 2022; Fanelli, 2018).

110 Reproducibility (*i.e.* different teams and experimental setups obtaining
111 similar results; Plesser, 2018) is one of the main criteria for evaluating robust
112 science and reliable conclusions. ~~The term "reproducibility" is a relative
113 concept and has known various definitions depending on field and context.
114 Reproducibility of analyses In ecological sciences, most in-situ observations
115 are not strictly reproducible due to stochasticity. Accordingly, the focus has
116 been directed towards the reproducibility of analyses ("computational
117 reproducibility") is defined by Cohen-Boulakia *et al.* (2017) as the ability of
118 distinct analyses to reach to the same conclusion. over the reproducibility of
119 data collection (Powers & Hampton, 2019; Samota & Davey, 2021).
120 Reproducibility can be achieved at different levels of the analytical workflow,
121 from primary data access to results. Archmiller *et al.*, 2020 and Minocher *et al.*,
122 2021 tried to evaluate computational reproducibility in 74 studies in
123 wildlife science and 560 studies in biological and behavioural sciences.
124 Although these authors found high rates of computational reproducibility
125 when data and analytical procedures could be fully retrieved, they
126 encountered significant difficulty in retrieving the data files and analytical
127 procedures in most studies.~~

128 ~~Given the high complexity and the massive amount of information
129 required to retrieve results using a broad range of data and methods,
130 achieving sufficient reproducibility must be facilitated. In addition,
131 researchers are increasingly challenged to stay up-to-date with the ever-
132 growing number of advanced methods and technologies for data acquisition,
133 storage, and analysis (Hampton *et al.*, 2017). Providing technical and
134 practical support to reduce the perceived complexity of analytical workflows
135 could increase and accelerate the diffusion of good practices in the research
136 community, fostering understanding for a wider audience thereby facilitating
137 transparency and improving reproducibility. Here, we explore how
138 computational reproducibility can be easily implemented in ecological
139 sciences using simple and practical guidelines.~~

140 In the current context of the global biodiversity crisis, the scientific
141 community needs to use all available data and provide as robust as possible
142 evidence regarding the state and dynamic of ecological systems, from
143 genetic to ecosystem. At the same time, using analytical tools to provide

144 robust evidence can be complex and may require advanced skills that are not
145 widely available across the scientific community ([Hampton et al., 2017](#)).
146 Therefore, operational solutions and methodological guidelines can allow ~~the~~
147 analytical workflows to be more accessible without degrading the scientific
148 quality of ~~the analyses~~, and thus, promote efficient and broad deployment of
149 ~~good practice~~ best practices.

150 Is the ecology community failing to meet ~~good practice~~ best practices?

151 The first step towards reproducibility is knowing current ~~good practice~~ best
152 practices and recommendations. Among them, the FAIR principles (Wilkinson
153 et al., 2016), for which the availability of the data and the code used for each
154 published result is an essential criterion, may be key for appropriate
155 management through the data life cycle (Michener, 2015). The FAIR
156 principles (see also CARE principles by Carroll et al., 2020) are considered as
157 a founding framework to share data along four important elements:
158 "Findable" for humans and machines; "Accessible" with a detailed access
159 procedure; "Interoperable" for interaction with other data or applications;
160 "Reusable" in an identical or different context. In addition to these principles,
161 propositions have been delimited within several thematic communities in
162 ecology to evaluate and enhance best practices application, notably the
163 Species Distribution Modelling communities (Araújo et al., 2019; Zurell et al.,
164 2020).

165 ~~In 2022, Gomes and collaborators identified 12 barriers to data and code~~
166 ~~sharing, ranging from unclarity of processes to fear of inappropriate use and~~
167 ~~insecurities around data and code quality (Gomes et al., 2022).~~ Although data
168 accessibility has been substantially improved in ecology during the past
169 decade, sharing analytical scripts and codes remain largely marginal (~~Ivimey-~~
170 ~~Cook et al., 2023~~ [Archmiller et al., 2020](#); ~~).~~ ~~According to~~ [Culina et al., 2020](#);
171 ~~Minocher et al., 2021; Ivimey-Cook et al., 2023).~~

172 ~~in a "random sample of 346 nonmolecular articles published between~~
173 ~~2015 and 2019", 79% had data availability but only 27% had code availability-~~
174 ~~despite a tendency for journals to encourage code sharing (75% of assessed~~
175 ~~ecological journals).~~


176 ~~Low code availability compared to data availability may suggest a lack of~~
177 ~~technical solutions for sharing computing codes. Nevertheless, many~~
178 ~~repositories dedicated to sharing code exist, such as GitHub~~
179 ~~(<https://github.org>), which software developers widely use to collaborate and~~
180 ~~share codes publicly and privately. Besides, the Software Heritage initiative~~
181 ~~automatically archives all openly available code from GitHub, ensuring long-~~
182 ~~term preservation (<https://archiveprogram.github.com>; Di Cosmo & Zacchiroli,~~
183 ~~2017). Alternatively, other solutions for data archiving may be used, even if~~
184 ~~not explicitly focused on code sharing (e.g., Zenodo, national public~~
185 ~~repositories; see also TRUST principles for data repositories, Lin et al., 2020).~~

186 However, even if ~~long-term public archiving of~~ sharing code is necessary to
187 achieve good computational reproducibility, it is insufficient. Therefore, ~~many~~
188 ~~guidelines and principles have been developed in the recent years. Among~~
189 ~~others~~, the utilisation of computational workflows has been suggested as a
190 solution for improving computational reproducibility (Cohen-Boulakia et al.,

191 2017; Grüning *et al.*, 2018) through software such as Snakemake (Köster &
192 Rahmann, 2012), Nextflow (Di Tommaso *et al.*, 2017), or Galaxy (The Galaxy
193 Community, 2022). A workflow is generally defined as a sequence of distinct
194 computational tasks for a particular objective (Goble *et al.*, 2020). As such, a
195 workflow represents the backbone of a single specific analysis. Throughout
196 the analytical procedure, a typical workflow starts with raw data, which can
197 be extracted from several databases or data files and processed through a
198 series of analytical steps. The products resulting from these analytical steps
199 (*i.e.* the outputs of the computational workflow) can be data files, graphic
200 representations and any associated metrics. ~~In this respect, computer code
201 can also be considered as research data (Borgman, 2020).~~

202 When properly designed, a certain level of reproducibility can be easily
203 achieved since workflow languages naturally capture the following four key
204 elements (Cohen-Boulakia *et al.*, 2017):

- 205 – the specificities of the workflow, the analysis steps and associated
206 tools;
- 207 – the workflow entries, datasets and parameters;
- 208 – the environment and context of the use of the workflow;
- 209 – the results obtained and the outputs of the workflow.

210 In the original publication of Wilkinson *et al.* (2016), the focus of FAIR
211 principles was mainly on observational data. However, the principles can be
212 applied to software and computational workflows (Lamprecht *et al.*, 2019;
213 Goble *et al.*, 2020). For instance, a code shared as supplementary material of
214 a non-open access publication could be considered as "Interoperable" but is
215 not easily "Findable", "Accessible", or "Reusable". In contrast, a large block of
216 code consisting of several hundred lines, from data pre-processing to final
217 results and graphics as pictured in the Graphical abstract , may require
218 efforts to understand and adapt to other kinds of data ("non-reusable"),
219 mainly if annotations or comments are limited. Similarly, an analytical
220 procedure shared without indicating the versions of hardware, software, and
221 packages has a low chance of producing identical outputs, making it non-less
222 reproducible. These issues may harm the scientific community by preventing
223 fully transparent communication among users about knowledge production
224 and practice comparison. They can also be detrimental to individual authors,
225 when they need to update or run new analyses.

226 Impact on Ecology Research

227 The efficiency of the expertise and research is greatly affected by the lack
228 of computational reproducibility and FAIRness of analytical procedures. FAIR
229 research data was estimated to save 10.2 billion € per year in Europe
230 (Munafò *et al.*, 2017; European commission, 2018⁹; Gomes *et al.*, 2022).
231 ~~Indeed, analyses and underlying conclusions cannot have a tangible impact if
232 the raw data, the analytical procedures, and the outputs resulting from these
233 procedures are not easily findable, accessible, interoperable and reusable.~~
234 Moreover, consistent application of reproducibility and FAIR principles will
235 improve trust in research studies and scientific reports (Powers & Hampton,
236 2019; Lortie, 2021; Jenkins *et al.*, 2023).

237 The widespread use of computational languages to process large-scale
238 data and analyse complex systems has been a major advance in studying the
239 ecosphere at any spatio-temporal scale (Michener & Jones, 2012; Farley *et al.*,
240 2018). ~~Even if computational capacity may represent a significant limitation
241 for analysing large data files or using resource-intensive algorithms (Green &
242 Figuerola, 2005), computation clusters nowadays exist to overcome such
243 challenges (Hampton *et al.*, 2017; Larcombe *et al.*, 2017).~~ However, the ever-
244 growing technical and programming skills required to take advantage of such
245 computational solutions by the scientific community raise new challenges
246 (Jetz *et al.*, 2019; Leroy, 2022; Boyd *et al.*, 2023). The use of increasingly
247 complex analytical solutions, paired with different approaches or
248 programming languages, mechanically reduces the number of potential users,
249 limiting collaboration and fragilising fundamental pillars of scientific
250 knowledge such as the peer-review process and critical evaluation. As a
251 response to this situation, adequate training was identified by life science
252 researchers (*Community Survey Report*, 2013; Williams & Teal, 2017;
253 Larcombe *et al.*, 2017), as it would help involve more people in the
254 understanding of current analytical solutions and benefit to scientific
255 cooperation (Touchon & McCoy, 2016; Gownaris *et al.*, 2022). Research is
256 typically structured through a highly competitive organisation, with a
257 potentially detrimental effect on scientific knowledge (Fang & Casadevall,
258 2015). Instead, fostering collaboration and collective intelligence by
259 promoting transparent sharing of analytical procedures, would offer more
260 persitent and robust ways to achieve actionable science (Ellemers, 2021).
261 Such efforts would be of paramount importance in environmental sciences
262 and the conservation of biodiversity by providing governance and guiding
263 actions with increasingly robust evidence (Keenan *et al.*, 2012).

264 Are there simple and ready-to-use solutions?

265 In this note, we aim to promote the reuse of existing concepts and
266 solutions as pillars toward better practices for ecological analyses by
267 providing a streamlined framework. We believe the atomisation-
268 generalisation framework presented in the second part of this note
269 represents an operational and actionable path for researchers and experts to
270 attain levels of ~~good practice~~ best practices (e.g. reproducibility, FAIR, open
271 science, R compendium; Casajus N., 2023) with no more investment than
272 they are able or willing to provide (Field *et al.*, 2014). Atomisation is used to
273 refer to the identification of single analytical steps constituting an analytical
274 procedure. It is a non-standard term introduced in this note to convey the
275 idea of analytical “atoms”. As for atom particles that etymologically
276 correspond to “indivisible” but are composed of subatomic particles, an
277 analytical atom represents a single analytical step composed of several
278 functions. Generalisation involves the alteration of an analytical step to
279 enlarge its applicability in diverse contexts and for diverse purposes.

280 This framework has been formalised while building the Galaxy-Ecology
281 (Galaxy-E) initiative (see section III). Galaxy (The Galaxy Community, 2022) is
282 a workflow-oriented web platform for sharing and processing ~~research~~ data. It
283 allows scientists to shareing, developing, and useing various datasets and

284 data processing tools (e.g. data formatting, statistical tests, graphic
285 representations). ~~Many scientific workflow management systems, such as~~
286 ~~Snakemake and Nextflow, operate from the command line. In ecology,~~
287 ~~numerous initiatives have tried to introduce such systems, starting with more~~
288 ~~user-friendly solutions. For example, the KNIME and Kepler systems with the~~
289 ~~CoESRA initiative (Collaborative Environment for Scholarly Research and~~
290 ~~Analysis) in Australia, or Taverna with the BioVeL initiative (Biodiversity~~
291 ~~Virtual e-Laboratory) in Europe. These systems are more accessible to new~~
292 ~~users by offering a graphical interface while achieving high specificity~~
293 ~~(Berthold et al., 2007; Hardisty et al., 2016). However, good computer~~
294 ~~programming or scientific workflow management knowledge is still necessary~~
295 ~~to use these applications correctly.~~

296 Galaxy is ready to use and has proved its efficiency and suitability in other
297 research fields, including genomics and climate science (Knijn et al. 2020;
298 Serrano-Solano et al., 2022). From a user's point of view, it offers extensive
299 computing power and a graphical interface to use analysis workflows, even
300 without experience in software development. Web-based access allows easy
301 sharing of analytical workflows between collaborators and with a broader
302 audience. Galaxy supports tools in almost any computational language,
303 including R and Python, two of the most used languages in ecology, with
304 many packages dedicated to ecological and biodiversity-oriented analyses
305 incorporated (Lai et al., 2019).

306 Galaxy enables good reproducibility for data exploration and analyses,
307 helps compute intricate analyses on big data files, enables collaboration, and
308 can support the teaching process. Galaxy-E is a Galaxy server dedicated to
309 ecological analyses maintained by the European Galaxy team (supported by
310 the German Federal Ministry of Education and Research and the German
311 Network for Bioinformatics Infrastructure), and is available at
312 <https://ecology.usegalaxy.eu>.

313 Galaxy-E is a demonstration platform for applying good practicebest
314 practices such as the FAIR principles and computational reproducibility for
315 analytical procedures in ecology. Hence, this technical note is partly Galaxy-
316 oriented, not to present the platform as a prescriptive solution but to give an
317 operational example of the good practicebest practices it helps to achieve.
318 ~~Recommendations described in this note regarding the construction of an~~
319 ~~analytical procedure on Galaxy are meant to be transposable to local code~~
320 ~~development or another consistent workflow engine.~~

321 Framework towards good practicebest practices

322 Atomisation: what is it and why?

323 Atomisation refers to dividing an analytical procedure into several
324 specific steps ("atoms"; Graphical abstract ②) generating a suite of
325 elementary analytical steps as pictured in the Graphical abstract ③. Breaking
326 down the analytical process into atoms functioning as building blocks allows
327 for better understanding, modularity, and visibility of the analytical flow. It
328 permits making it more accessible to a broader audience or facilitating the
329 peer-review process. Indeed, an extended one-block code that imports raw

330 data, makes pre-processing steps (e.g. filter, formatting), conducts analyses
331 (e.g. distribution study, modelling), and performs final representations of
332 results (e.g. maps, plots) can be challenging to understand and reuse by
333 others or even the same person after some time.

334 McIntire *et al.* (2022) described the PERFICT approach (Prediction,
335 Evaluation, Reusability, Free access, Interoperability, Continuous workflows,
336 and routine Tests) to set a new foundation for models in predictive ecology.
337 This can be applied more generally to the analytical procedure in ecology and
338 biodiversity. In their article, McIntire and collaborators make an analogy
339 between code development and Lego® construction, similar to our definition
340 of atomisation. Functions are a workflow's most fundamental analytical steps
341 and can be seen as modular pieces, alike single pieces of Lego®. Modules
342 can be created from a single or series of successive functions comparably as
343 in Lego® structures made of several pieces (e.g. meant to build cars, houses,
344 or road). These modules (or atoms, tools) can be used as standalone or
345 combined to make simple to complex analytical workflows ~~such as~~(e.g. data
346 formatting or curation, running statistical models, or generating graphical
347 elements for visualisation). Doing so, the atomisation approach may facilitate
348 sharing or teaching analytical practices since beginners can easily
349 understand the general organisation of the analytical procedure by simply
350 reading the list of steps in the analysis with a limited degree of complexity.
351 Decoupling programming skills from analytical skills can make data
352 processing more accessible to a wider audience. Indeed, once each
353 elementary step is clearly identified and delimited along the atomisation
354 process, it is easier to grasp the whole analytical procedure and focus on the
355 review of each step at a time or (re)use it. New workflows can further be
356 generated by recombining existing, validated or peer-reviewed elementary
357 steps in innovative ways. This process can save time, increase confidence,
358 and avoid potential programming mistakes, allowing greater focus on
359 understanding the analytical workflow.

360 Generalisation: what is it and why?

361 Generalisation ~~is~~ refers to the modification of an analytical procedure to
362 make it applicable to many settings, by removing specificities related to a
363 particular data file or data format. Generalisation aims to optimise the
364 reusability at different times (e.g. regular result update), enlarge the
365 application of a given analysis to different input data files while keeping the
366 initial analytical procedure fully reproducible as pictured in the Graphical
367 abstract 4. Generalising an analytical step requires identifying key steps and
368 invariant parameters from those that must be adaptable to allow for the
369 analysis to be applied to specific characteristics of various datasets. These
370 parameters must be implemented to be easily modified if needed.
371 Generalisation can be tricky because the higher the flexibility of an analytical
372 step, the greater the risk of errors in its use. This is why generalisation should
373 be complemented by clear statement and an implementation of red flags and
374 warnings to prevent such events. As with atomisation, generalisation is
375 primarily a conceptual way to build analytical procedures. It requires minor

376 change of practices to reach certain degree of generalisation, avoiding
 377 additional effort later on for reusability, reproducibility, and share.

378 How to do atomisation and generalisation with computer codes: Finding
 379 balance

380 Breaking down codes into elementary steps to achieve atomisation is not
 381 an intuitive task at first as it may target a single function or a more intricate
 382 set of several functions. There could be different degrees of atomisation,
 383 depending on the grain required to decompose the analytical process (fig. 1;
 384 tab. 1). The application of general guidelines and good practice/best practices
 385 implies finding a balance between the most appropriate degree of
 386 atomisation and generalisation. This depends on the type of analytical
 387 procedure or the targeted audience (e.g. with different interests and
 388 programming skills). Attention to this balance is critical to ensure that the
 389 analytical procedures could be reused. For instance, a workflow in which each
 390 function would be considered as a unique elementary step would optimise
 391 the flexibility but may likely add unnecessary complexity. At the other
 392 extreme, considering a whole analytical workflow as an elementary step may
 393 make it ready-to-use and simplify its application, but would be too coarse and
 394 therefore limit flexibility by violating the principle of atomisation.

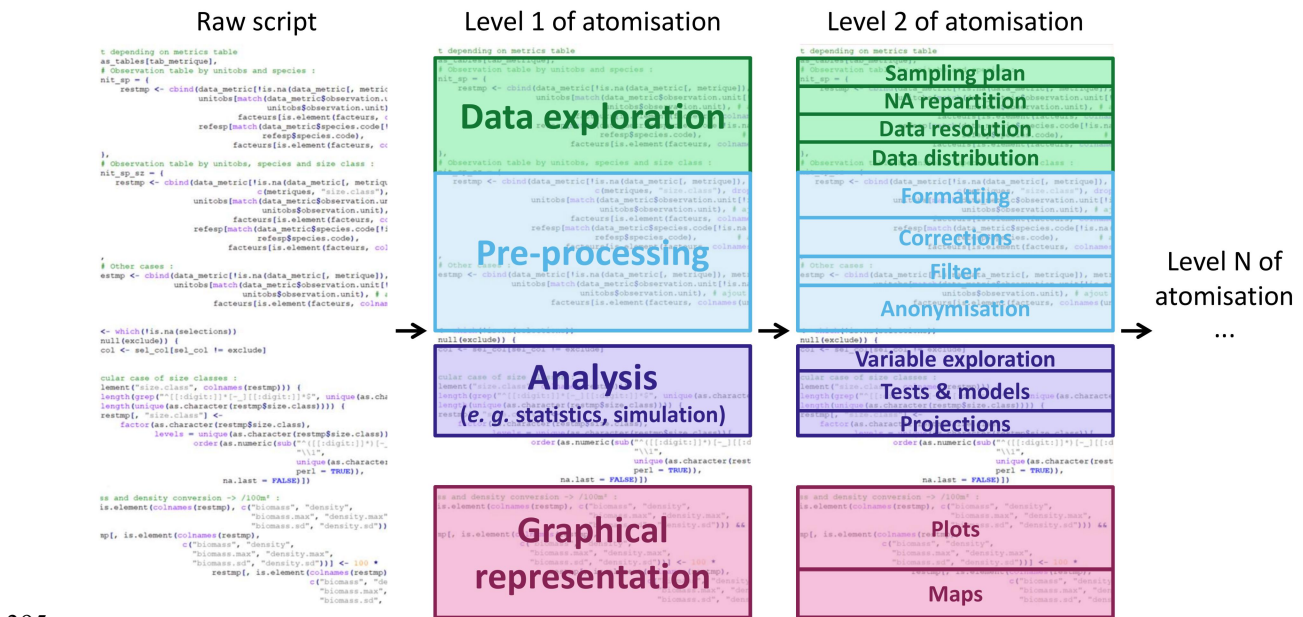


Figure 1 - Illustration of the atomisation of an existing code

Table 1 - Example of atomisation levels

Level 1 - big shape	Level 2	Level 3
Data exploration	Sampling plan	Complete Balanced
	Missing values	Proportion Distribution
	Data granularity	Geographic resolution Temporal resolution Measure resolution
	Data distribution	Geographic coverage Temporal coverage Measures ranges Summaries
...
Pre-processing	Formatting	Change file format Change general format
	Corrections	Remove special characters Remove low trust observations Correct measures
	Filtering	Remove unwanted observations
	Anonymisation	Anonymise names Anonymise localities Anonymise species
...
Analysis	Variable exploration	PCA Collinearity Correlation
	Unimodal tests	Linear Models χ^2 Student
	Statistical models	Generalised Linear Models Generalised Additive Models Random Forest
	Models Evaluation	Evaluation metrics (e.g. AIC, Jaccard) Validation methods
	Projections	Geographical projections Temporal projections
...
Representation	Plot	Raw variables Modelled results
	Map	Observations Projections
...

398 **A fF**ew changes in code-writing habits can enhance the reusability of the
399 analytical procedure by generating easy-to-understand analytical procedure
400 without investing much time. It is best to develop each elementary step
401 directly in separate code files and to give details of the order in which
402 elementary steps are used for each analytical workflow. To ensure
403 reproducibility and traceability of the results, each computation of the
404 analytical workflow should be associated with the details of the parameters
405 settings and datasets used. From a practical point of view, a couple of
406 recommendations could be made for coding elementary steps in order to
407 facilitate generalisation and ease the reuse. Once each elementary step is
408 defined, we recommend all dependencies (e.g. software version, packages,
409 libraries and their versions) to be set at the same place, at the start of the
410 code, followed by modular parameters (e.g. input file location and name,
411 column selection, modelling parameters, data specificities, output saving
412 location). When the script of the elementary step is completed, modular
413 parameters should be the only part of the code that may be modified in
414 future reuse. Dependencies and subsequent computational tasks should be

415 left untouched to ensure the integrity of the analysis and then, reproducibility.
416 In the end, it is best to add an open-source license to any analytical
417 procedure shared publicly (e.g. MIT, GPL). It permits to clearly state the terms
418 and conditions of diffusion, share and reuse.

419 As such, atomisation and generalisation may overcome social or
420 psychological barriers related to transparent sharing, either related to
421 securing ownership (e.g. DOI) and to embarrassment or fear during a peer-
422 review process (Gomes *et al.*, 2022).

423 Atomisation and generalisation are related and complementary concepts.
424 Atomisation into adequate elementary steps is necessary to properly
425 generalise an analytical procedure as it permits to enhance the modularity of
426 the procedure and its capacity to be tailored to different data types.
427 Atomisation and generalisation must be applied from the earliest stages of
428 the programming development of any analytical procedure in order to
429 achieve:

430 – Greater transparency, even for beginners, since the relevance and
431 coherence of each step and their successive arrangement along the
432 analytical procedure should be appraised independently of the
433 programming skills;

434 – Time savings;

435 – Greater reusability;

436 Modularity of the elementary steps, to rearrange them differently if
437 needed.

438 –

439 Entering a new dimension: the Galaxy-E initiative example

440 Developing open and properly atomised and generalised analytical
441 procedures can already represent a significant step forward in terms of ~~good~~
442 practicebest practice. Galaxy is a good illustration of atomisation and
443 generalisation with easier management of analytical workflows. The platform
444 proposes many analytical tools that represent generalised and atomised
445 elementary steps. These tools are modular and openly licensed, which
446 permits to build generalised workflows as pictured in the Graphical abstract
447 5.

448 Galaxy-E is mostly aimed at scientists that process biodiversity data and
449 already have an understanding of the general functioning of the analytical
450 procedures they want to produce. The rationale for a user would be to create
451 or reuse analytical workflows with high FAIRness in a collaborative and open
452 source platform. It can be used for individual analyses as well as for
453 collaborative projects. In some cases, if the analytical procedure is already
454 clearly defined, it can be used by citizens or for teaching.

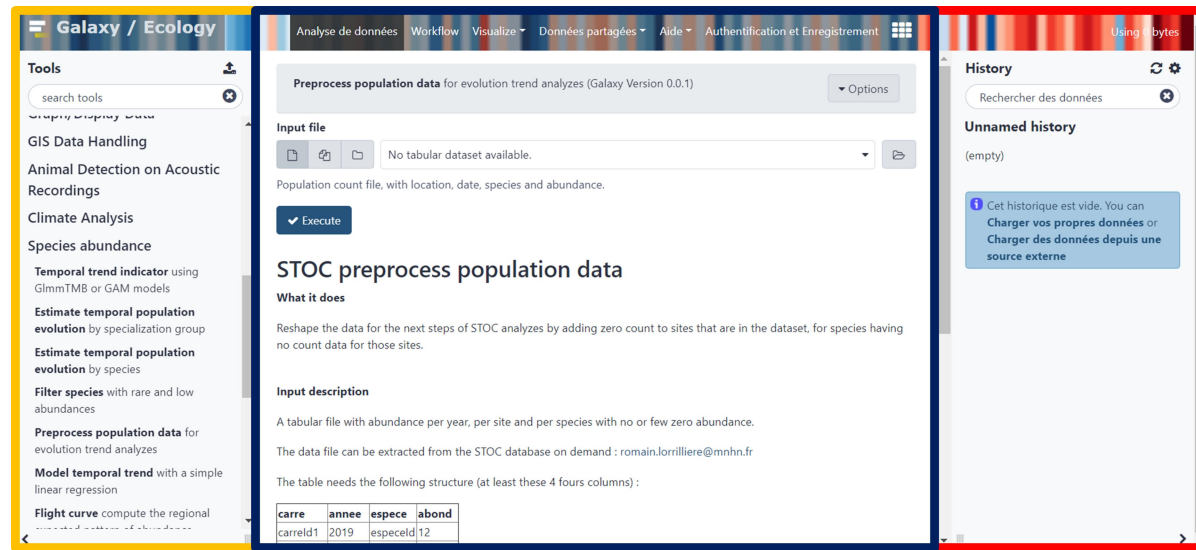
455 It benefits from the same advantages as the framework presented in the
456 previous section and can help achieve a further level of FAIRness as a
457 demonstration platform to package analyses in an accessible and user-
458 friendly manner (tab. 2).

459 Galaxy as a demonstration platform to package analyses in an accessible and user-friendly manner can help achieve
460 a further level of FAIRness. Any analytical procedure can be adapted on the platform and Galaxy can be used through-
461 the whole data life cycle (https://rdmkit.elixir-europe.org/galaxy_assembly). Throughout this note, many ways to
462 contribute to Galaxy are discussed in their conceptual and methodological aspects. One can use off-the-shelf tools,
463 workflows, and tutorials to design an analytical procedure, or suggest, develop, and share new workflows and tutorials,
464 two aspects that do not require coding skills. Eventually, one can modify or develop entirely new tools with any
465 computational language to make them accessible to all users on any Galaxy server. The Galaxy platform emphasises (i)-
466 accessibility of tools and data even without programming experience, (ii) reproducibility through the easy creation and
467 reuse of analysis workflows, (iii) transparency through the open-source distribution of underlying codes; and (iv)-
468 community support. In 2022, Gomes and collaborators identified 12 barriers to data and code sharing, ranging from
469 unclarity of processes to fear of inappropriate use and insecurities around data and code quality (Gomes *et al.*, 2022).

470 There are different Galaxy servers, at global, continental, and national levels (European and French levels for
471 example), but also according to the fields (e.g., biomedical, ecology, climate). The Galaxy-E initiative is hosted by
472 European (<https://ecology.usegalaxy.eu>) and French (<https://ecology.usegalaxy.fr>) servers.

473 Datasets can be uploaded on a Galaxy server from a local device, an online server, or a database. Users can then
474 access every available tools (fig. 2, left panel) to modify, explore, and analyse their data. All tools used, parameters,
475 and data (inputs and outputs) of the analysis are saved in a private “Galaxy history” (fig. 2, right panel), documenting
476 every step of the analytical procedure and recording the provenance of each output. From any history, the user can
477 extract a workflow (fig. 3) or directly share or publish the history itself.

478



479

480

Figure 2 – Galaxy-Ecology users' interface <https://ecology.usegalaxy.eu>. Yellow panel on the left: analysis tool list; blue panel in the middle: current tool interface; red panel on the right: Galaxy analysis history

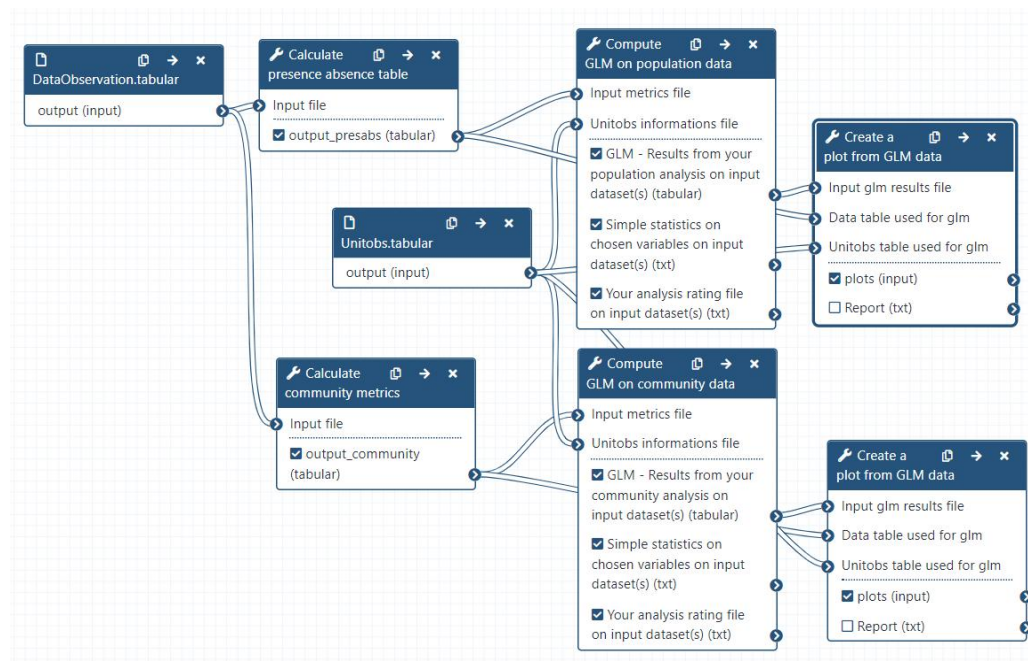


Figure 3 - Representation of a Galaxy workflow in the editing interface of a Galaxy server. Each box represents an analysis tool, and the lines represent the flow of data through the tools

Anyone can use the tools on Galaxy and/or develop new tools and workflows to make them available to all by publishing them in the shared Galaxy ToolShed (<https://toolshed.g2.bx.psu.edu/>) which ensures that the tools and dependencies can be installed on any Galaxy servers. By definition, a Galaxy workflow already has a degree of atomisation (each tool represents an elementary step) and generalisation and benefits from the same advantages as the framework presented in the previous section in good practices (tab. 2). Any analysis history or workflow can be shared and enriched in parallel by several users, facilitating teamwork. Galaxy-Ecology has implemented workflows for biodiversity data exploration, eDNA processing, general population and community metrics and models, ecoregionalisation, NDVI (Normalised difference vegetation index) computation with Sentinel-2 data among others (see some examples: <https://workflowhub.eu/workflows/657>) and tutorials for several of them are available on the GTN platform (see <https://training.galaxyproject.org/training-material/topics/ecology>).

494 ~~Recommendations described in this note regarding the construction of an analytical procedure on Galaxy are meant~~
495 ~~to be transposable to local code development or another consistent workflow engine.~~

Table 2 - Comparison between the atomisation-generalisation framework and Galaxy for the achievement of ~~good~~
~~practice~~best practices. Limitations are occasionally raised with short advice to mitigate them when relevant

		Atomised-generalised code	Galaxy
Reproducibility and transparency	Environment, software and package versions	Can be indicated but possibly hard to manage Can also be set as an output of the analysis (e.g. session info) Packages written in each coded elementary step or using a versioning system such as Conda	Entirely packaged with Conda package manager and BioContainers Possibility to store analytical procedures as containers for persistent execution
	Inputs and parameters	One must keep track of different parametrisation and input settings at each computation	Automatically tracked and shareable with the “Galaxy history”
	Peer-review	Organisation of the analytical procedure reviewable by non-code developers Code developers might be able to detect errors as it is easier in shorter scripts Transparency over the development process achievable through Git	Reviewable “Galaxy history” and re-executable workflow <u>Continuous</u> <u>p</u> Peer-reviewed <u>of</u> tools with open-source code Transparency over the development process through Git The workflows can be reviewed by the Intergalactic Workflow Commission (IWC) for best practices
	Output provenance	Can be tracked and reproduced in some cases	Tracked with the “Galaxy history” and reproducible with workflow
FAIR principles	Findable	If properly shared	Web-based solution Unified system for data and software citation and attribution Tools can be made available on several servers Tools can be linked to tools registries and annotated with different ontologies Annotated workflows findable on WorkflowHub (https://workflowhub.eu) and Dockstore (https://dockstore.org)
	Accessible	If properly shared	Free distribution of tools via the Galaxy ToolShed and workflows via WorkflowHub and Dockstore under an open-source licence
	Interoperable	When properly generalised, different elementary steps should be useable in interaction with each other	Use different software, computational language and library versions on a single platform with the Conda package management system Workflows exportable in JSON and shareable through several standards (e.g. Common Workflow Language; Crusoe <i>et al.</i> , 2022 and Research Object Crate; Soiland-Reyes <i>et al.</i> , 2022)
	Reusable	Generalised elementary steps are reusable and adaptable with different analytical procedure, parametrisation and/or inputs	Tools, histories and workflows are re-executable, reusable and adaptable with different analytical procedure, parametrisation and/or inputs. Open-source code can be used outside of a Galaxy server
Technical and knowledge gaps	Understandability	The analytical procedure is clearer when properly atomised	Tools interface, workflow annotations, help sections and tutorials are a valuable help
	Teaching opportunities	Learning the analytical procedure design separately from computing languages, giving structure to trainees Reusability of elementary steps for trainees	Experimenting with intricate analyses without computer code first Tutorials and videos from Galaxy Training Network (https://training.galaxyproject.org) Galaxy community
	Computing capacity	Need for a computation cluster if large data or demanding algorithm	HPC (High Performance Computing) through an interface Bulk (meta)data manipulation
Collaboration and attribution	Analysis design and development	Achievable through collaborative code-editing applications	With anyone through a Galaxy server
	Citation	Easy reuse of openly shared elementary steps could lead to higher citation rates	Each tool, workflow, and tutorial are provided with a unique identifier for proper attribution and citation

499 The Galaxy platform emphasises (i) accessibility of tools and data even
500 without programming experience, (ii) reproducibility through the easy
501 creation and reuse of analysis workflows, (iii) transparency through the open-
502 source distribution of underlying codes; and (iv) community support.

503 Galaxy is ready to use and has proved its efficiency and suitability in other
504 research fields, including genomics and climate science (Knijn *et al.* 2020;
505 Serrano-Solano *et al.*, 2022). For scientists, from a user's point of view, it
506 offers extensive computing power and a graphical interface to use analysis
507 workflows, even without experience in software development. Web-based
508 access allows easy sharing of analytical workflows between collaborators and
509 with a broader audience. Galaxy supports tools in almost any computational
510 language, including R and Python, two of the most used languages in ecology,
511 with many packages dedicated to ecological and biodiversity-oriented
512 analyses incorporated (Lai *et al.*, 2019).

513 Anyone can use the tools on Galaxy and/or develop new tools and
514 workflows to make them available to all by publishing them in the shared
515 Galaxy ToolShed (<https://toolshed.g2.bx.psu.edu/>) which ensures that the
516 tools and dependencies can be installed on any Galaxy servers. Any
517 analytical procedure or workflow can be shared and enriched in parallel by
518 several users, facilitating teamwork.

519 Galaxy is a powerful platform enabling researchers to readily move
520 towards best practices. The Galaxy interface mitigates the difficulties
521 associated with library management and code development, which permits
522 simpler access to complex analytical methods. One can focus on the analysis
523 itself and its concepts, rather than on syntax difficulties or cluster
524 programming, disconnecting the study of data analysis concepts from the
525 study of computing languages.

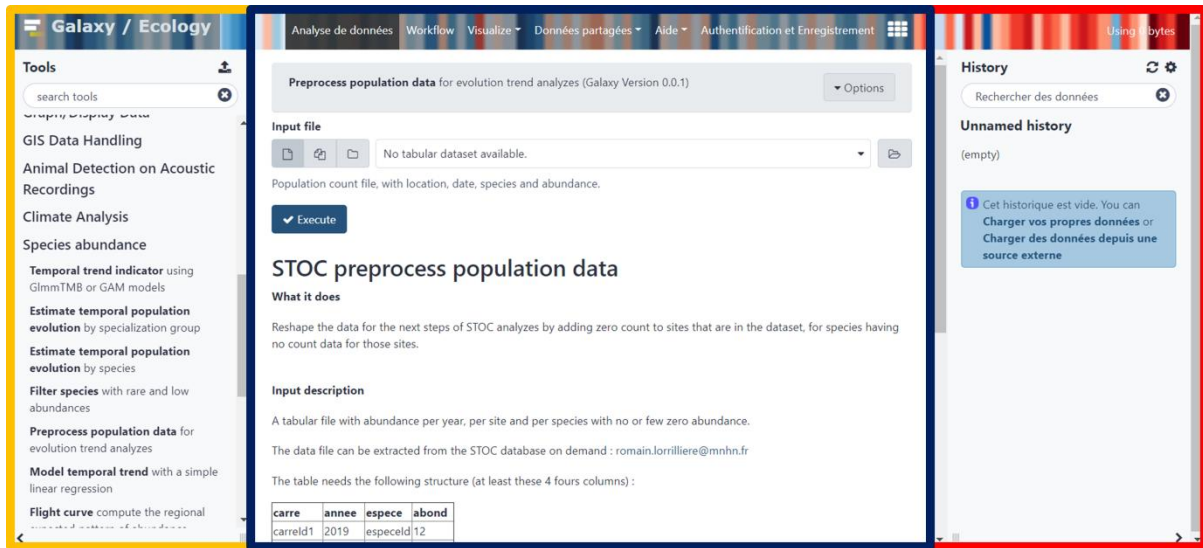
526 The platform is community-driven which permits continuous peer review of
527 the platform and of the tools, workflows and tutorials provided. Many tutorials
528 are available on the Galaxy Training Network (GTN) which is a valuable asset
529 to the accessibility and reusability of tools and workflows (Batut *et al.*, 2018;
530 Hiltemann *et al.*, 2023).

531 If enough researchers and experts start using and contributing to the
532 platform, the number and content of available analytical procedures could
533 expand at the same pace as latest analytical methodologies are integrated to
534 research processes. If a different platform fits best and is more widely used
535 by ecological and biodiversity scientific communities in the end, the work
536 done on Galaxy will not be lost as tools are easily transposable to other
537 interfaces (e.g. scripts directly usable with R, Python, etc., translation of
538 workflows to other workflow engines).

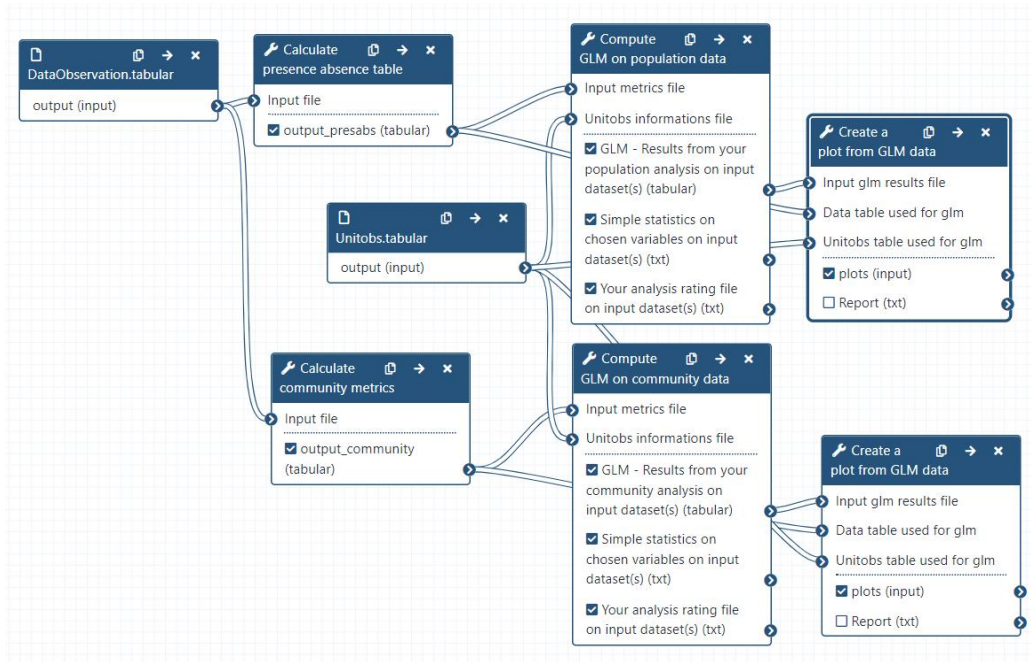
539 There are different Galaxy servers, at global, continental, and national
540 levels (European and French levels for example), but also according to the
541 fields (e.g., biomedical, ecology, climate). The Galaxy-E initiative is hosted by
542 European (<https://ecology.usegalaxy.eu>) and French
543 (<https://ecology.usegalaxy.fr>) servers.

544
545 Datasets can be uploaded on a Galaxy server from a local device, an
546 online server, or a database. Users can then access every available tool (fig.

547 2, left panel) to modify, explore, and analyse their data. All tools used,
 548 parameters, and data (inputs and outputs) of the analysis are saved in a
 549 private “Galaxy history” (fig. 2, right panel), documenting every step of the
 550 analytical procedure and recording the provenance of each output. From any
 551 history, the user can extract a workflow (fig. 3) or directly share or publish
 552 the history itself. Workflows are reusable through WorkflowHub
 553 (<https://workflowhub.eu>) or Dockstore (<https://dockstore.org>) and exportable
 554 in CWL and RO-CRATE standards.



555
 556 **Figure 2 - Galaxy-Ecology users' interface** <https://ecology.usegalaxy.eu>.
 557 Yellow panel on the left: analysis tool list; blue panel in the middle:
 558 current tool interface; red panel on the right: Galaxy analysis history



559
 560 **Figure 3 - Representation of a Galaxy workflow in the editing interface**
 561 of a Galaxy server. Each box represents an analysis tool, and the lines
 562 represent the flow of data through the tools

563 Any analytical procedure can be adapted on the platform and Galaxy can
564 be used through the whole data life cycle ([https://rdmkit.elixir-](https://rdmkit.elixir-europe.org/galaxy_assembly)
565 [europe.org/galaxy_assembly](https://rdmkit.elixir-europe.org/galaxy_assembly)). One can use off-the-shelf tools, workflows, and
566 tutorials to design an analytical procedure, or suggest, develop, and share
567 new workflows and tutorials, two aspects that do not require coding skills.

568 Galaxy-Ecology has implemented workflows for biodiversity data
569 exploration, eDNA processing, general population and community metrics
570 and models, ecoregionalisation, NDVI (Normalised difference vegetation
571 index) computation with Sentinel-2 data among others (see some examples:
572 <https://workflowhub.eu/workflows/657>) and tutorials for several of them are
573 available on the GTN platform (see [https://training.galaxyproject.org/training-](https://training.galaxyproject.org/training-material/topics/ecology)
574 [material/topics/ecology](https://training.galaxyproject.org/training-material/topics/ecology)).

575 Eventually, one can modify or develop entirely new tools and workflows
576 with any computational language to make them accessible to all users on any
577 Galaxy server.

578 Galaxy is an utterly participative platform and several ways to participate
579 to Galaxy exist depending on one's skills, available time, and needs. Anyone
580 can participate to the Galaxy-Ecology initiative by notably:

- 581 - [Sharing datasets, histories and workflows;](#)
- 582 - [Giving feedback on servers, tools, and workflows;](#)
- 583 - [Sharing tools and workflows ideas \(eventually with code\) through Git](#)
584 [issues;](#)
- 585 - [Asking for tool modifications through issues;](#)
- 586 - [Modifying existing tools or proposing new tools through GitHub or](#)
587 [GitLab;](#)
- 588 - [Writing or contributing to a GTN tutorial on a specific functionality or a](#)
589 [workflow on the Galaxy Training Network platform;](#)
- 590 - [Create learning pathways, a set of tutorials curated by community](#)
591 [experts to form a coherent set of lessons around a topic, building up](#)
592 [knowledge](#) ([https://training.galaxyproject.org/training-](https://training.galaxyproject.org/training-material/learning-pathways)
593 [material/learning-pathways](https://training.galaxyproject.org/training-material/learning-pathways));
- 594 - [Propose training events and help users in the utilisation of a workflow](#)
595 [and tutorial.](#)

596
597 Analyses are rarely computed only once. Any analysis with a
598 generalisation potential is a suitable candidate to be Galaxy-fied. A
599 methodological framework is presented in online supplementary material
600 ([https://github.com/ColineRoyaux/Galaxy_Templates/blob/main/Methods/Methods](https://github.com/ColineRoyaux/Galaxy_Templates/blob/main/Methods/Methods%20-%20How%20to%20Galaxy-fy%20your%20analytical%20procedure_.md)
601 [ods%20-%20How%20to%20Galaxy-](https://github.com/ColineRoyaux/Galaxy_Templates/blob/main/Methods/Methods%20-%20How%20to%20Galaxy-fy%20your%20analytical%20procedure_.md)
602 [fy%20your%20analytical%20procedure_.md](https://github.com/ColineRoyaux/Galaxy_Templates/blob/main/Methods/Methods%20-%20How%20to%20Galaxy-fy%20your%20analytical%20procedure_.md)) at three levels depending on
603 potential interests, computing language skills, and willingness to invest more
604 or less time in the process: (i) 'user' relying on existing Galaxy tools and
605 workflows to analyse data (lower time investment), (ii) 'developer' relying on
606 existing and validated analytical procedure to develop Galaxy tools and
607 workflows (highest time investment), and (iii) 'trainer' relying on existing
608 Galaxy tools to share workflows and create training material (variable time

609 ~~investment). The 12 barriers to data and code sharing raised by Gomes et al.,~~
610 ~~(2022) can be at least partially addressed by Galaxy (see fig. S1).~~

611 ~~Galaxy is a powerful platform enabling researchers to readily move towards~~
612 ~~good practices. The Galaxy interface mitigates the difficulties associated with~~
613 ~~library management and code development, which permits simpler access to~~
614 ~~complex analytical methods. One can focus on the analysis itself and its~~
615 ~~concepts, rather than on syntax difficulties or cluster programming,~~
616 ~~disconnecting the study of data analysis concepts from the study of~~
617 ~~computing languages.~~

618 ~~The Galaxy Training Network (GTN) is a valuable asset to the accessibility~~
619 ~~and reusability of tools and workflows (Batut et al., 2018; Hiltemann et al.,~~
620 ~~2023). The Galaxy Training platform (<https://training.galaxyproject.org>) is an~~
621 ~~open, FAIR, collaborative platform compiling a variety of tutorials written by~~
622 ~~researchers, administrators, developers, and other contributors. These~~
623 ~~tutorials not only aim to teach how to use Galaxy, and take advantage of~~
624 ~~advanced features such as Interactive Tools (i.e. interactive applications~~
625 ~~within Galaxy, e.g. Windows desktop, Rstudio, R Shiny apps), but also how to~~
626 ~~run and interpret scientific analyses through detailed step-by-step guides.~~

627 Discussion and limitations

628 Levels of good practice

629 As highlighted in previous sections, there are many ~~good practice~~**best**
630 ~~practices~~ and recommendations existing for analytical procedures, data
631 management, and computational code development. The levels of
632 application of these ~~good practice~~**best practice** fall within a continuum
633 offering many possibilities. From the lowest to the highest ~~good practice~~**best**
634 ~~practice~~ levels for a published work there can be for example:

- 635 – Raw data and analytical procedure are not shared, only processed and
636 interpreted results along with a brief description of methods.
- 637 – Pre-processed data is shared, and methods are described in the word-
638 limit given by the publisher (example: tables of metrics and how it was
639 calculated).
- 640 – Raw data and source code are shared on a repository. Software and
641 package versions are not specified and there is no guaranty to be able
642 to reproduce the analytical procedure.
- 643 – Raw data and atomised - generalised source codes are shared on a
644 repository with specified hardware, software and dependencies
645 versions. Input parameters are recorded in an attached file.
- 646 – Raw data is shared with proper metadata and an actionable version of
647 the whole analytical procedure is traceable, ready to use and
648 eventually reuse on other data types. Such level can be attained
649 notably using Galaxy.
- 650 – All results and conclusions are published as an executable paper with
651 analyses and workflows implemented and executable directly in the
652 shared article (Strijkers et al., 2011).

653 Executable Papers (Strijkers *et al.*, 2011) can require significant time and
654 resource investment as well as good knowledge of programming languages,
655 making it an admirable but hard-to-attain goal.

656 Atomisation and generalisation of computer codes can represent a
657 relatively low investment strategy to attain certain levels of best practices
658 such as transparency and reusability. It also carries advantages such as
659 easier peer review, modularity of analytical procedures and, consequently,
660 time savings. Indeed, applying the framework is not sufficient to attain the
661 highest levels of best practices. For reproducibility and transparency, the
662 management of the environment, softwares and package versions can be
663 hard to maintain and record. A comprehensive tracking of input, outputs and
664 codes requires meticulous management of files arborescence in the
665 environment. Additionally, non-code developers will be able to partially
666 review the analytical procedure only if the workflow is clearly outlined in an
667 adapted format (e.g. table, graphical representation). Accessibility and
668 findability of the atomised and generalised analytical procedure is dependent
669 of its proper sharing (e.g. persistent link, open repository).

670 ~~On Galaxy can represent an easier gateway towards higher levels of best~~
671 ~~practice, as any available tool can be easy to use. Sharing sharing a complete,~~
672 ~~detailed and (re-)executable analytical procedure is facilitated as through~~
673 ~~provenance is tracked and metadata is automatically metadata~~
674 ~~enrichmentd. Finally, a Galaxy history or workflow can be made accessible~~
675 ~~to anyone (See methods section for details on the use of Galaxy). In~~
676 ~~comparison, many scientific workflow management systems, such as~~
677 ~~Snakemake, Nextflow or the R package Targets, operate from the command~~
678 ~~line. In ecology, numerous initiatives have tried to introduce such systems,~~
679 ~~starting with more user-friendly solutions. For example, the KNIME and Kepler~~
680 ~~systems with the CoESRA initiative (Collaborative Environment for Scholarly~~
681 ~~Research and Analysis) in Australia; Taverna with the BioVeL initiative~~
682 ~~(Biodiversity Virtual e-Laboratory) in Europe; or very recently, the BON in a~~
683 ~~Box pipeline engine. These systems are more accessible to new users by~~
684 ~~offering a graphical interface while achieving high specificity (Berthold *et al.*,~~
685 ~~2007; Hardisty *et al.*, 2016; <https://boninbox.geobon.org/>).~~
686 ~~However, good computer programming or scientific workflow management knowledge is still~~
687 ~~necessary to use these applications correctly.~~

688 In comparison to the atomisation-generalisation framework, Galaxy can be
689 rightfully seen as heavier for experienced programmers as it requires to learn
690 to use a new platform. Additionally, mMore effort may be required on Galaxy
691 when an additional analytical step needs to be developed, but the Galaxy
692 community can be an efficient crutch on which hard-pressed scientists can
693 rely. Indeed, one can ask for help on the implementation of tools whether one
694 knows computing languages and can share their code or not.

695

696 ~~A deeply collaborative initiative-~~

697 ~~Galaxy is an utterly participative platform. Any analysis history or workflow~~
698 ~~can be shared and enriched in parallel by several users, facilitating teamwork.~~
699 ~~As discussed earlier, several ways to participate to Galaxy exist depending on-~~

700 ~~one's skills, available time, and needs. In the methods section, three ways to~~
701 ~~participate to Galaxy are distinguished: "as a user", "as a developer" and "as~~
702 ~~a trainer". One is not confined to only one of these roles; this distinction is~~
703 ~~more of a handy way to give structure to the methodology depending on~~
704 ~~one's skills, available time and needs. Anyone can participate to the Galaxy-~~
705 ~~Ecology initiative by notably:-~~

706 ~~Sharing datasets, histories and workflows;~~

707 ~~Giving feedback on servers, tools, and workflows;~~

708 ~~Sharing tools and workflows ideas (eventually with code) through Git~~
709 ~~issues;~~

710 ~~Asking for tool modifications through issues;~~

711 ~~Modifying existing tools or proposing new tools through GitHub or GitLab;~~

712 ~~Writing or contributing to a GTN tutorial on a specific functionality or a~~
713 ~~workflow on the Galaxy Training Network platform;~~

714 ~~Create learning pathways, a set of tutorials curated by community experts~~
715 ~~to form a coherent set of lessons around a topic, building up knowledge~~
716 ~~(<https://training.galaxyproject.org/training-material/learning-pathways/>);~~

717 ~~Propose training events and help users in the utilisation of a workflow and~~
718 ~~tutorial.~~

719 ~~Galaxy is community driven which permits continuous peer review of the~~
720 ~~platform and of the tools, workflows and tutorials provided. If enough~~
721 ~~researchers and experts start using and contributing to the platform, the~~
722 ~~number and content of available analytical procedures could expand at the~~
723 ~~same pace as latest analytical methodologies are integrated to research~~
724 ~~processes. If a different platform fits best and is more widely used by~~
725 ~~ecological and biodiversity scientific communities in the end, the work done~~
726 ~~on Galaxy will not be lost as tools are easily transposable to other interfaces~~
727 ~~(e.g. scripts directly usable with R, Python, etc., translation of workflows to~~
728 ~~other workflow engines), histories shareable as files and workflows reusable~~
729 ~~through [WorkflowHub](https://workflowhub.eu/) (<https://workflowhub.eu/>) or [Dockstore](https://dockstore.org)~~
730 ~~(<https://dockstore.org>) and exportable in CWL and RO-CRATE standards.~~

731 ~~Galaxy Ecology has implemented workflows for biodiversity data~~
732 ~~exploration, eDNA processing, general population and community metrics~~
733 ~~and models, ecoregionalisation, NDVI (Normalised difference vegetation~~
734 ~~index) computation with Sentinel-2 data among others (see some examples:~~
735 ~~<https://workflowhub.eu/workflows/657>) and tutorials for several of them are~~
736 ~~available on the GTN platform (see [https://training.galaxyproject.org/training-](https://training.galaxyproject.org/training-material/topics/ecology/)~~
737 ~~[material/topics/ecology/](https://training.galaxyproject.org/training-material/topics/ecology/)).~~

739 Conclusion

740 This ~~article~~ note showcases a simple proposition to achieve ~~good~~
741 ~~practice~~ best practices in analytical procedures with two plain guidelines:
742 atomisation and generalisation. This straightforward framework represents a
743 different manner to think and build analytical procedures; it doesn't require
744 using a new technology or learning to use a new software. In terms of
745 attaining higher levels of best practice, whether it is through the atomisation-

746 generalisation framework, Galaxy, a combination of the two or otherwise, the
747 optimal approach is to be determined by individuals depending on their
748 interests, projects, and available resources. Relying on existing solutions as
749 much as possible is, in our perspective, an efficient way to achieve a better
750 understanding of good practicebest practices and their implications. Given
751 the current environmental crisis, science has the major political and social
752 responsibility to maintain good levels of transparency, reproducibility and
753 efficiency.

754 Methods - How to Galaxy-fy your analytical procedure?

755 ~~Analyses are rarely computed only once. Any analysis with a~~
756 ~~generalisation potential is a suitable candidate to be Galaxy-fied. This~~
757 ~~methodological framework is presented at three levels depending on~~
758 ~~potential interests, computing language skills, and willingness to invest more~~
759 ~~or less time in the process: (i) 'user' relying on existing Galaxy tools and~~
760 ~~workflows to analyse data (lower time investment), (ii) 'developer' relying on~~
761 ~~existing and validated analytical procedure to develop Galaxy tools and~~
762 ~~workflows (highest time investment), and (iii) 'trainer' relying on existing~~
763 ~~Galaxy tools to share workflows and create training material (variable time~~
764 ~~investment). Of course, learning to use a new platform and trying to look~~
765 ~~differently at analyses is time-consuming in the short term, but saves time in~~
766 ~~the long run. Even if in the end the analysis is not made available on Galaxy,~~
767 ~~the work is not lost as each step helps the analysis to reach a higher level of~~
768 ~~good practice.~~

769 Guidelines "as a user"

770 ~~Whether one wants to design a new analysis directly on Galaxy or has~~
771 ~~already an established analytical procedure and wants to adapt it on Galaxy~~
772 ~~to make it easier to review and reuse, the following steps are approximately~~
773 ~~the same. As Galaxy already is a workflow-oriented platform with atomisation~~
774 ~~of steps, "atoms" of the analysis are apparent while building the analysis on~~
775 ~~Galaxy.~~

776 ~~The Galaxy platform offers many options that can be explored using the~~
777 ~~guided tours of the interface (on the welcome page or tab "Help - Interactive~~
778 ~~Tours"). Several tutorials are also available on the Galaxy Training Network~~
779 ~~(<https://training.galaxyproject.org>) to learn how to use Galaxy (e.g. topics~~
780 ~~"Introduction to Galaxy Analyses", "Using Galaxy and Managing your Data").~~
781 ~~Main steps of the implementation of an analytical procedure on Galaxy as a~~
782 ~~user are represented on figure 4.~~

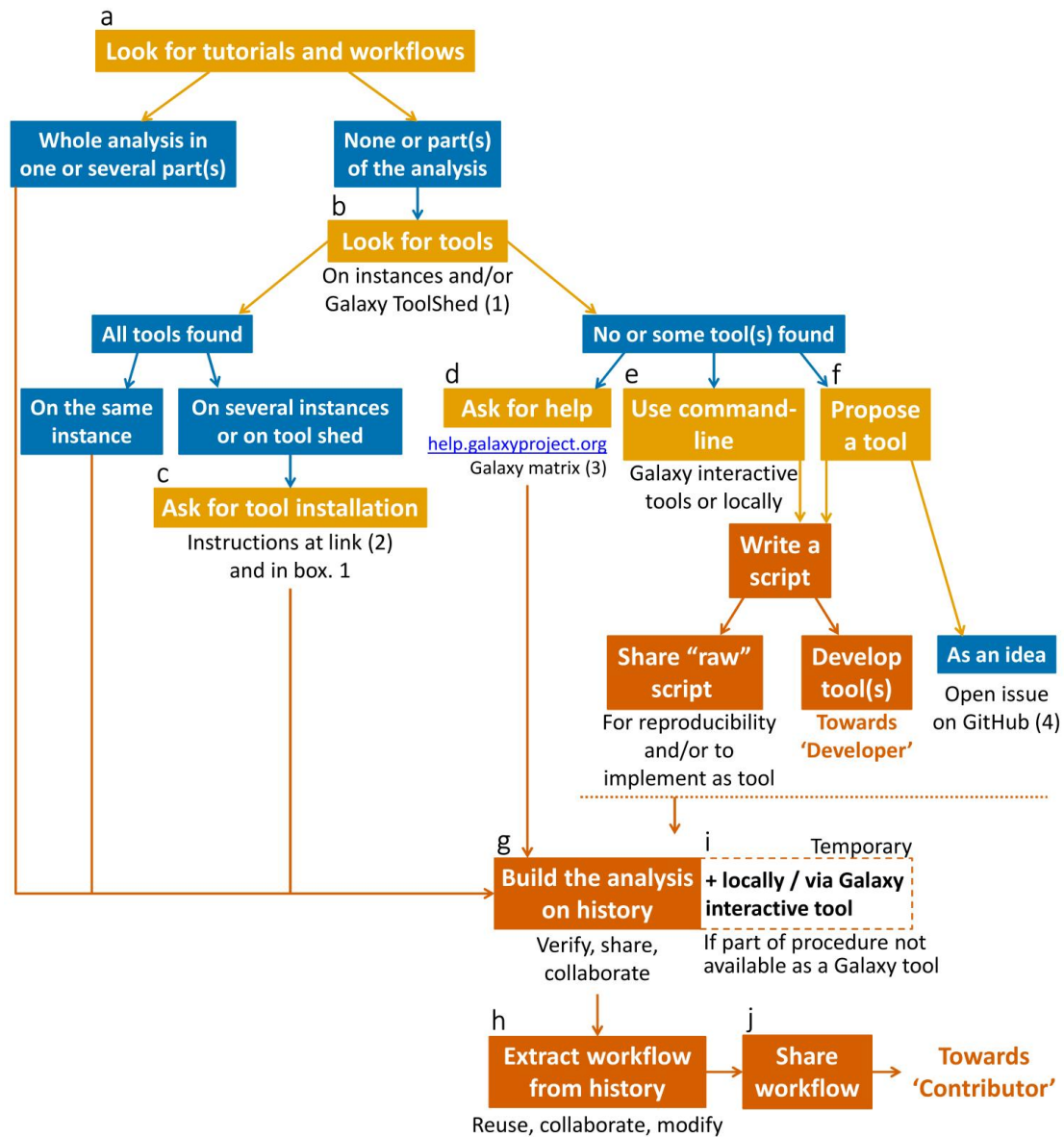


Figure 4— Decision tree and framework for Galaxy users relying on existing tools and workflows. The orange boxes represent actions. The blue boxes represent possible situations one may encounter during the procedure. The red boxes represent steps where one could stop, share the work, and then attain better reproducibility and FAIRness. Letters at the top left of boxes indicate which paragraph it refers to in the text. Links: (1) <https://toolshed.g2.bx.psu.edu> (2) <https://usegalaxy-eu.github.io/posts/2020/08/22/three-steps-to-galaxyfy-your-tool> (3) <https://matrix.to> (4) <https://github.com/galaxyecology>

(a) The first thing to do when starting an analysis on Galaxy is to look for tutorials on the Galaxy Training platform to benefit from others' experience. One tutorial may be enough to set the tracks for the whole analytical procedure, but it is also possible to use sub-parts of tutorials and/or associate several tutorials to complete steps of the procedure. Numerous ready-to-use workflows are also available on the Galaxy servers (tab "Shared Data Workflows") or could be imported from WorkflowHub or Dockstore, one may find one or several workflows to complete its analysis. High quality peer-

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801 reviewed Galaxy workflows are reported by the Intergalactic Workflow
802 Commission (IWC, <https://github.com/galaxyproject/iwc>). Additionally, it is
803 possible to seek for help by asking on the Matrix channel
804 (<https://gitter.im/Galaxy-Training-Network/Lobby>) or by opening a topic on
805 the Galaxy Help (<https://help.galaxyproject.org>).

806 (b) If the whole analytical procedure has not been fully covered with
807 available tutorials and workflows, almost 10,000 tools are available on the
808 Galaxy Tool Shed (<https://toolshed.g2.bx.psu.edu>) to connect the dots.

809 (c) One or several helpful tools might not be installed on the used Galaxy
810 server and one may need to ask for an installation (See box. 1 Ask for tool
811 installation).

812 **Box 1** Ask for tool installation. See [https://usegalaxy-](https://usegalaxy-eu.github.io/posts/2020/08/22/three-steps-to-galaxyfy-your-tool/)
813 [eu.github.io/posts/2020/08/22/three-steps-to-galaxyfy-your-tool/](https://usegalaxy-eu.github.io/posts/2020/08/22/three-steps-to-galaxyfy-your-tool/) for
814 more details

Fork: Act of creating a copy of a repository in one's personal space.

Commit: Act of submitting a modification to a file

Pull Request (PR): Act of proposing one or several Commit(s) to be integrated

Merge: Act of accepting the PR and integrate the modification proposed on the repository

Galaxy tools installation process is accessible to anyone, it is often explained directly in the "Read me" file on the server tools-
repository (usually on GitHub or GitLab). To ask for the installation of a tool one must:

Look for the tool repository on the Galaxy Tool Shed;

Look for the domain tools repository (e.g. <https://github.com/usegalaxy-eu/usegalaxy-eu-tools> for all Galaxy Europe servers;
<https://gitlab.com/ifb-elixirfr/usegalaxy-fr/tools> for Galaxy France);

Fork this repository and look for the .yaml file corresponding to the used server (e.g. ecology.yaml for the
<https://ecology.usegalaxy.eu> and <https://ecology.usegalaxy.fr> servers);

In the .yaml file, make a Commit to add the following lines with the name and owner of the tool (written on the tool repository
on the Galaxy Tool Shed) along with a suggested tool panel section in which the tool can be sorted:-

```
...  
name: pampa_presabs  
owner: ecology  
tool_panel_section_label: 'Species abundance'  
...;
```

PR the modification on the domain tools repository and wait for server maintainers' approval (merge) and/or suggestions. The
installation of tools might be rejected if the peer review process or relevance of the proposed tool is not adequate in the
server maintainers' opinion.

815 If there are still gaps in the analytical procedure that none of the existing
816 tools can fill, several options are available:

817 (d) Ask for help (see end of bullet a).

818 (e) Temporarily fill the gap with a command-line code locally or through a
819 Galaxy Interactive Tool (e.g. Rstudio, Jupyter notebook and Ubuntu desktop
820 interactive tools). The code can be shared or not.

821 (f) Propose a new tool by sharing the idea through a GitHub issue
822 (<https://github.com/galaxyecology>; preferably along with a code if existing).
823 Details on the task aimed and awaited input and output (i.e. full
824 specifications) of the tool along with references are of great help for potential
825 developers who may take over tool development. If one wants to try tool
826 development, see section 'As a developer'.

827 (g) Through these steps of looking for tutorials, workflows, and tools, the
828 analytical procedure is progressively designed on the Galaxy history. As each
829 Galaxy tool, parametrisation and provenance of each file produced is tracked
830 in the Galaxy history, one can try several tools with different parameters to
831 compare and find out which configuration seems the best. The Galaxy history

832 can be shared to anyone through a link to collaborate on the analysis or in a
833 peer-review process.

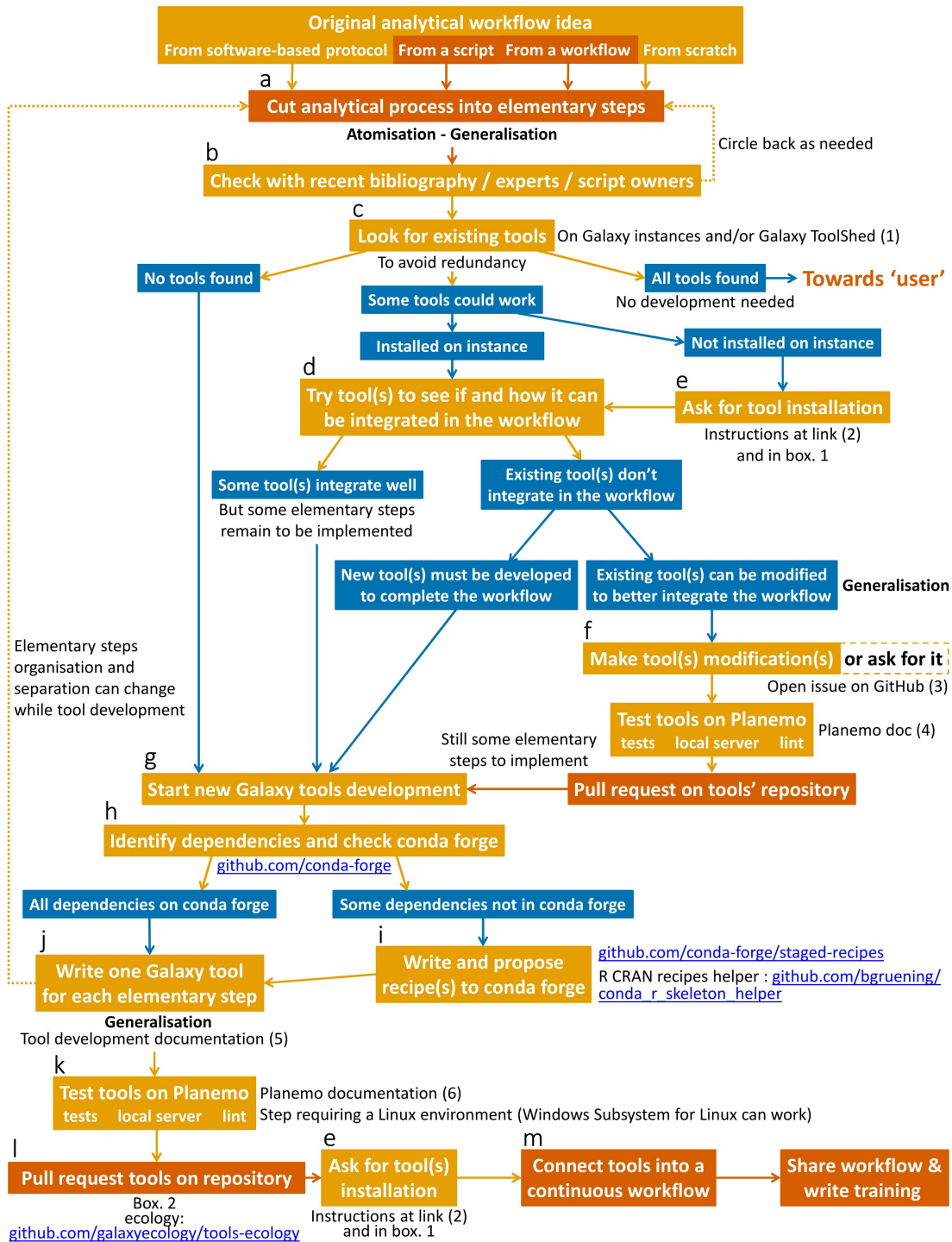
834 (h) When parametrisation stage is done and the analytical procedure is
835 complete, one can extract a workflow to reuse the analytical procedure on
836 new datasets.

837 (i) In the case of a missing tool and part of the analytical procedure is
838 temporarily performed outside Galaxy, one can build separate workflows,
839 between which data is downloaded to make required steps locally. A better
840 temporary solution is to program the launch of Galaxy Interactive Tools (e.g.
841 Posit (R), Jupyter notebooks, and Ubuntu desktop interactive tools) in the
842 workflow to keep most of the procedure on Galaxy. In this case, provenance
843 tracking can be secured partially by saving created objects, command history
844 (e.g. Rhistory), and running environment for example.

845 (j) Extracted workflow(s) can be shared with others for feedback or
846 collaboration, but it can also be shared publicly on Galaxy server(s) and/or
847 integrated to an article. When starting to share openly workflow(s), one is a
848 Galaxy contributor as well as a user (see section “As a trainer”).

849 Guidelines “as a developer”

850 Developing Galaxy tools requires time investment, especially at the
851 beginning to understand how Galaxy works and the architecture of the tools.
852 The development procedure can vary depending on the origin of the
853 analytical workflow idea which can be (i) existing code, a package, or a
854 workflow implemented elsewhere, (ii) an idea from a user proposal, (iii) a
855 published article or a personal need, and even (iv) an analytical procedure
856 using originally several interfaced tools. When an analytical procedure was
857 originally designed with atomisation and generalisation of elementary steps
858 in mind, the process of developing Galaxy tools should take a lot less time.
859 Main steps of the implementation of an analytical procedure on Galaxy as a
860 developer are represented on figure 5.



861

862
863
864
865
866
867
868
869
870
871
872
873

Figure 5— Decision tree and framework for Galaxy developers. Orange boxes represent actions, blue boxes represent possible situations one may encounter during the process and red boxes represent shareable steps where one could stop and still attain better reproducibility and FAIRness. Letters at the top left of boxes indicate which paragraph it refers to in the text.
 Links: (1) <https://toolshed.g2.bx.psu.edu> (2) <https://usegalaxy-eu.github.io/posts/2020/08/22/three-steps-to-galaxyfy-your-tool> (3) <https://github.com/galaxyecology> (4) <https://planemo.readthedocs.io/en/latest/index.html> (5) <https://docs.galaxyproject.org/en/latest/dev/schema.html> (6) <https://planemo.readthedocs.io/en/latest/index.html>

874 (a) The atomisation process starts at early stage of the design of an
875 analytical workflow before writing any computer code. Atomisation into
876 elementary steps provides clarity to the development phases. Ultimately, one
877 elementary step equals one Galaxy tool and the modular parameters
878 identified in the code for generalisation would be those that appear on the
879 tool interface.

880 (b) One can start by splitting essential steps of the analysis (e.g. pre-
881 processing, analyses, representations) and detailing each elementary step
882 afterward to get different atomisation resolutions (tab. 1; fig. 1). The first
883 atomisation is not a permanent choice and will certainly be refined over the
884 course of the development process. It is mainly useful as a medium for
885 researchers and other scientists to give feedback on the projected
886 architecture of the workflow and to have an overview of the analytical
887 procedure. As for any analysis, one must check if potential issues or red flags
888 were raised by the community on the methods used and take it into account
889 in the architecture of the workflow. At this point, any products generated
890 from the atomisation process can be shared and be useful to the scientific
891 community. For example, sharing a written description or a schematic
892 representation of the steps and organisation of an analytical procedure
893 (coded or not) is a valuable help for anyone trying to make a similar analysis.

894 (c) As a user would do and before starting tool development, one must
895 look for existing tools on Galaxy servers and Galaxy ToolShed
896 (<https://toolshed.g2.bx.psu.edu>) to avoid redundancy. If all needed tools are
897 available, one can directly build their workflow on Galaxy, see 'As a user'
898 section. Many tools are available on Galaxy for data manipulation. If one
899 needs a particular format or type of data there is high probability that it can
900 already be handled on Galaxy.

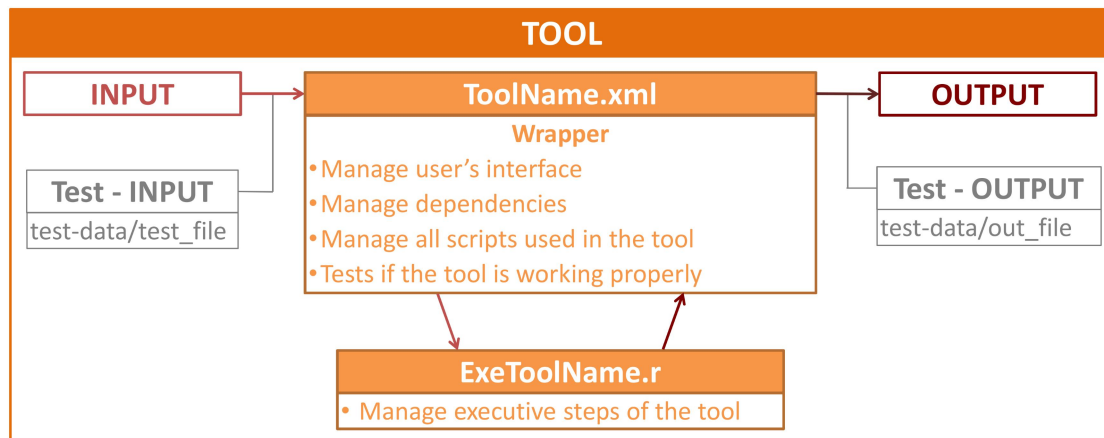
901 (d) If some tools could work in the workflow, one must test it to see if and
902 how it can be integrated.

903 (e) In the case some tools are not installed on the Galaxy server, ask for
904 tool installation (see box.1)

905 (f) Selected tools might not integrate precisely as aimed, if the input or
906 the output is not formatted as projected in the primary workflow design,
907 other tools added before and/or after might solve the problem. If such tools
908 are not available or the problem is more about a missing parameter or
909 methodology, it might be more coherent to modify existing tool(s) than
910 developing entirely new ones. One can open a new GitHub issue to ask for
911 modifications on the tool repository (found on the Galaxy ToolShed) or
912 directly suggest modifications on the tool. When modifying a tool, the process
913 is approximately the same as for developing an entirely new tool (explained
914 in the next paragraph) only the Pull Request for modifications should be
915 opened on the tool repository.

916 (g) The Galaxy community has made available a lot of documentation
917 resources for tool development on the GTN Training platform (category
918 "Development in Galaxy"; [https://training.galaxyproject.org/training-
919 material/topics/dev](https://training.galaxyproject.org/training-material/topics/dev)) and on the General Galaxy documentation
920 (<https://docs.galaxyproject.org>;
921 <https://docs.galaxyproject.org/en/latest/dev/schema.html>).

922 Galaxy tools have a common architecture (fig. 6). Each tool consists of an
 923 XML (Extensible Markup Language) wrapper which defines input file(s) and
 924 parameters that are presented to the end-user in the Galaxy web interface
 925 ("ToolName.xml" in fig. 6). Inputs provided through the interface can be
 926 processed with code in any computing language ("ExeToolName.r" in fig. 6).
 927 Outputs of the code are also specified in the XML file and are made available
 928 to the user in the Galaxy history at the end of the computation.



929
 930 **Figure 6** – Schematic representation of the simplified architecture of an
 931 example Galaxy tool using R language. From the input files and
 932 parameters provided by the user, the tool will launch an analytical
 933 procedure through the XML and R files to produce the outputs.

934 At least one unit test is mandatory to make sure a tool works and
 935 produces the expected outputs. This also facilitates maintenance, as tests
 936 will indicate if the functionality is preserved after tool updates. To do so, the
 937 test is written in the XML file with all parameter settings, input and expected
 938 output files (stored in a sub-directory "test-data") or characteristics of the
 939 expected output.

940 This organisation can be more elaborate, especially when developing
 941 several tools at the same time. For example, parts of XML files may repeat
 942 themselves in the different tools and one can create a supplementary XML
 943 file to write this repeating part once as a macro and call ('expand') it as
 944 needed, which saves time and space. The same type of repeating patterns
 945 can occur in the computing code and one should create a functions file to
 946 avoid copy-pasting of many lines in several separate code files.

947 Detailed documentation of the XML wrapper files is available in Galaxy,
 948 see <https://docs.galaxyproject.org/en/master/dev/schema.html>, as well as
 949 tutorials (<https://gxy.io/GTN:T00117>). An empty Galaxy tool template in R
 950 language is available in the following repository:
 951 https://github.com/ColineRoyaux/Galaxy_Templates/tree/main/R_Tool_template.

952
 953 (h) To begin development, it is best to have knowledge of the required
 954 informatics dependencies of the tool(s) such as software versions, packages
 955 and their versions to directly check their availability on Conda Forge
 956 (<https://conda-forge.org/feedstock-outputs>).

957 (i) Some dependencies might not be available, and, in this case, one must
958 write and propose a recipe to the Conda Forge on GitHub
959 (<https://github.com/conda-forge>), for guidelines see [https://conda-](https://conda-forge.org/#add_recipe)
960 [forge.org/#add_recipe](https://conda-forge.org/#add_recipe). For Python and R packages available on Pypi or CRAN
961 respectively, helper codes are available to automatically generate recipes,
962 see <https://github.com/conda/grayskull> and
963 https://github.com/bgruening/conda_r_skeleton_helper (by B. Grüning),
964 respectively. Dependencies of the Galaxy tools are called in the XML file.

965 (j) Generalisation of computational code is especially important while
966 developing the Galaxy tool to make sure the tool is useful to the largest
967 audience. It is difficult to think about all possible purposes of a tool, one will
968 likely miss some aspects but as Galaxy is a participative platform, anyone
969 can ask for modifications or make it themselves. The format of the input file
970 is a critical aspect of developing a Galaxy tool, while other aspects of the
971 format can be left to the users' choice or imposed. For example, on Galaxy,
972 the preferred format for table input is tab-separated values (TSV or "tabular").
973 Many tools on Galaxy are available to convert file formats (e.g. from CSV to
974 tabular).

975 For example, a typical choice to make as a developer when developing a
976 tool dealing with tables is to ask the user to specify through the interface
977 which column contains a specific variable, or to require a column name to be
978 present in the input file for the tool to find the variable. The first option is
979 more generalised as it is easier for the user to select a column directly on the
980 interface rather than change column names in the data files. The second
981 option can however be chosen when the tool uses a lot of columns in
982 different input tables or has a lot of intricate parameters to avoid
983 unnecessary complexity of the tool interface. This option can also be
984 consistent for tools using input data file written in a standardised way, as
985 Darwin-core data standard for example.

986 Depending on the type of manipulations and analyses made in the tool,
987 many parameters might be useful for users to customise such as the type of
988 model, the distribution law of the data, the corrections to make on the data,
989 the level of resolution or the type and format of output(s). Prior discussions
990 on the workflow with experts and researchers on the analytical procedure can
991 permit to raise important parameters for the users to set. Another good way
992 to get a view on what kind of parameters can be useful for users is to check
993 directly for parameters in the functions used in the computational code and
994 identify which ones are important for the computation and might be critical
995 for users to set. These parameters can be provided with default values if the
996 user does not provide a custom value. An "advanced parameters" collapsible
997 section can also be implemented to keep the interface simple while still
998 permitting flexibility for experimented users. Finally, to check if a workflow is
999 properly generalised, one can seek input files of different origins from open
1000 data repositories or ask scientists to test their tools.

1001 It is impossible to prevent all possible misuses of software and such events
1002 occur also when using command-line functions. Implementation of error and
1003 warning messages in the computing code is the best way to avoid misuse
1004 (e.g. wrong input format or parameter selection). One can also use the

1005 interface, the help section of tools, and training to help users to set
1006 parameters properly and raise red flags on the use of tools and workflows
1007 (e.g. the tool cannot be used on some types of data, types of modelling
1008 interact badly with some parameters settings or data distributions). If
1009 possible, implementing verification steps in the tools to give feedback to the
1010 user on how the computation went is also a good way for the user to get
1011 hindsight on the results (e.g. quantity of data that couldn't be used in the tool,
1012 models' evaluation variables, summary plots).

1013 (k) To verify tools syntax (lint), run unitary tests (test), and deploy a local
1014 Galaxy server to test tools interface (serve), one must use Planemo, the
1015 Galaxy Software Development Kit (Bray *et al.*, 2023). Planemo is a command-
1016 line tool used on a Linux environment (see documentation
1017 <https://planemo.readthedocs.io/en/latest>. For Windows users, Planemo can
1018 work on a WSL (Windows Subsystem for Linux) or using cloud development
1019 environment like GitPod. Galaxy Tool development can take many forms; the
1020 computational code can be developed beforehand on the local environment
1021 or, together with the XML file and be tested directly through a local interface
1022 deployed for testing. Each strategy has different pros and cons depending on
1023 the type of analytical procedure, the origin of the workflow, and the
1024 developer personal preference and knowledge.

1025 (l) When ready, tool(s) can be proposed to a collaborative Galaxy tool
1026 repository (for ecology: <https://github.com/galaxyecology/tools-ecology>; see
1027 box. 2 for procedure on GitHub) for peer-review by the community.

1028 **Box 2** Definitions of Git terminology and procedure for proposing a
1029 tool to a Galaxy repository

Fork: Act of creating a copy of a repository in one's personal space

Commit: Act of submitting a modification to a file

Pull Request (PR): Act of proposing one or several Commit(s) to be integrated

Merge: Act of accepting the PR and integrate the modification proposed on the repository

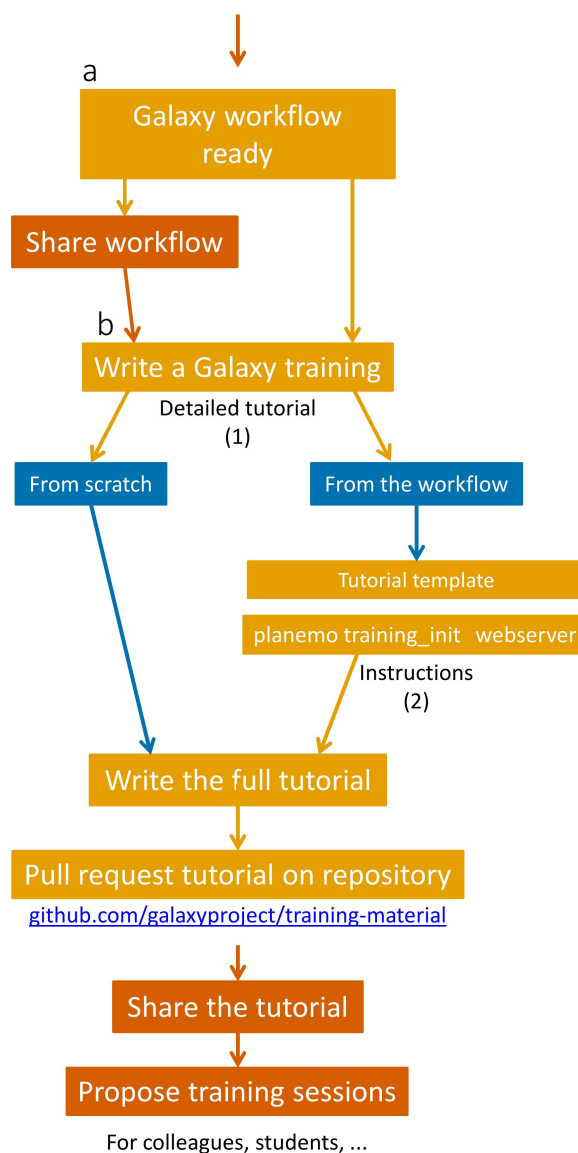
One has to fork the repository to add their new tool with a Commit and propose a PR against the original Galaxy repository with a brief description of the aims of developed tool(s) (PR example: <https://github.com/galaxyecology/tools-ecology/pull/50>). When a PR is opened on the repository, verification ("Check jobs") of the tool(s) compatibility, syntax, development good practices and proper running are made automatically. If there are problems, one can check output logs of what went badly and try to correct it while scientists invested in the Galaxy community give feedback on the tool(s). When checks are finally passed and code is peer-reviewed by the community, the PR is merged and the tool(s) made available on the Galaxy ToolShed within a few days. One may then ask for tool installation on any server (see box. 1 Ask for tool installation).

1030 (m) Once all developed tools are available on the Galaxy server, one can
1031 build a workflow as a user would do, share it and eventually write a training
1032 on the use of the workflow, see section "as a trainer".

1033 Guidelines "as a trainer"

1034 Main steps of the implementation of an analytical procedure on Galaxy as
1035 a trainer are represented on figure 7.

From 'User' or 'Developer'



1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

Figure 7—Decision tree and framework for Galaxy trainers. Orange boxes represent actions, blue boxes represent possible situations one may encounter during the process and red boxes represent shareable steps where one could stop and still attain better reproducibility and FAIRness. Letters at the top left of boxes indicate which paragraph it refers to in the text.

Links: (1) <https://training.galaxyproject.org/training-material/topics/contributing/tutorials/create-new-tutorial/tutorial.html> (2) <https://training.galaxyproject.org/training-material/topics/contributing/tutorials/create-new-tutorial/tutorial.html#create-the-skeleton-of-the-tutorial>

1048

1049

1050

1051

1052

(a) When an analytical procedure is built on Galaxy, one can extract a workflow from the history created. This workflow can be modified afterward to add annotations, comments, and flags. To make their workflow more generalised, one can leave parameters empty and users will have to set these parameters each time the workflow is launched. This workflow can be

1053 shared to contribute to Galaxy. Ultimately, it could be submitted to IWC and
 1054 be made available on WorkflowHub and/or Dockstore.

1055 (b) Eventually, one can write a tutorial on the GTN or a blog post on the
 1056 Galaxy Community Hub to get better visibility and broadcast valuable
 1057 elements on the use of the workflow. GTN tutorials are written in markdown.
 1058 One can start from scratch, but it is easier to start from a template generated
 1059 from an existing Galaxy workflow using the dedicated webserver
 1060 (<https://ptdk.apps.galaxyproject.eu>) or the command-line software Planemo
 1061 (documentation: <https://planemo.readthedocs.io/en/latest>). Indeed, this
 1062 approach only requires adding any needed explanations between the auto-
 1063 generate “hands-on” boxes containing tools and parameters instructions.
 1064 Many tutorials explain the different ways to contribute to the GTN (e.g.
 1065 tutorials, slides, videos, training sessions, quizzes) in the contributing topic
 1066 on the GTN: [https://training.galaxyproject.org/training-](https://training.galaxyproject.org/training-material/topics/contributing)
 1067 [material/topics/contributing](https://training.galaxyproject.org/training-material/topics/contributing) Introduction on the creation of a new hands-on
 1068 tutorial is detailed in this tutorial: [https://training.galaxyproject.org/training-](https://training.galaxyproject.org/training-material/topics/contributing/tutorials/create-new-tutorial/tutorial.html)
 1069 [material/topics/contributing/tutorials/create-new-tutorial/tutorial.html](https://training.galaxyproject.org/training-material/topics/contributing/tutorials/create-new-tutorial/tutorial.html). Like
 1070 tools, contributions to Galaxy Training Material are proposed through GitHub
 1071 (<https://github.com/galaxyproject/training-material>). Available tutorials are
 1072 publicly and freely available and can be openly shared to colleagues and
 1073 students and be used during courses and training sessions.

1074 Appendices

1075 **Table S1** – Barriers and solutions to data and code-sharing raised by
 1076 Gomes *et al.* (2022), along with corresponding solutions on the Galaxy
 1077 platform.

Barriers	Solutions and arguments from Gomes <i>et al.</i> (2022)	How Galaxy addresses the barrier
Unclear-sharing-process	Use FAIR principles Try, even if it is not perfect Look for online resources Ask editorial support staff and institutional libraries	FAIR and workflow-oriented platform Easy sharing of computational procedures (“Galaxy history” and/or workflow) as a link or a file attached to a publication Available online resources and forums for help
Complex-workflows	Process and clean data with reproducible code Detailed description of data processing steps Use non-proprietary files or softwares Avoid manual tasks	Reproducible workflows and visualisation of analytical procedure with the interface (fig. 3) “Galaxy history” tracks provenance of outputs and details of the data processing steps Possibility to add annotations and write a tutorial Open source platform Manual tasks can be recorded in workflows
Large-data-files	Free cloud storage Bundle smaller datasets	Free cloud storage (storage extension on demand) and High-Performance Computing
Insecurity	Share to trusted peers and/or on pre-prints servers before formal peer review Review before publication ensures in higher quality results Foster an inclusive environment promoting growth over criticism and shame “Perfect code” doesn’t exist	“Galaxy history” and workflow record the whole analytical procedure, it is private by default and can be shared to specific users or through a link making review by trusted peers easier and faster before public sharing Peer-reviewed tools
Unclear-value	Uncertainty about potential reuse should not present a barrier to sharing	Sharing an analytical procedure is not only relevant for others’ reuse but also for collaboration, peer review, and teaching. Sharing tools or workflows with Galaxy enables overcoming this uncertainty Methods of the note aims to facilitate this process and ensure it is properly made, adding a layer of clarity regarding the value of shared codes
Inappropriate-use	Metadata information with thorough description of datasets and processes, terms and consideration of	Raise major red flags or potential misuse in the help section and/or in the tool execution by validating input before tool

	reuse and any limitations, assumptions, caveats, and shortcomings Include contact information	execution: Implemented errors and warnings in the code to prevent directly prohibitive use of tools. Write execution suggestions and guidelines in the workflow annotations and/or associated tutorial. Possibility to produce editable report when executing a workflow or from the "Galaxy history"
Rights	Use open repositories instead of attaching code and data directly to the article as supplementary material Use data and code licenses Seek for help with institutional libraries and offices dedicated to copyright, open science and commercialisation	Open-source platform and tools shared through public servers prevents copyright issues Each Galaxy tool related code must have a license. Annotation of workflows with license Use of GitHub (or GitLab) to share code and workflows
Sensitive content	Aggregating, generalising or anonymising data	Sharing data and analytical procedure is up to the user Available tool to anonymise geographical coordinates on Galaxy
Transient storage	Archive data in permanent repositories Avoid proprietary files (e.g. Microsoft suite files) Use tools to promote backwards compatibility and portability of softwares and packages within different operating systems (e.g. containers, Jupyter notebooks)	Use of Software Heritage through GitHub to archive code Promotes non-proprietary files (e.g. TSV, fasta) Version-controlled tools to ensure the consistency and persistence of analyses even over updates Conda package manager and BioContainers to ensure cross-operating system compatibility for any programming language Containerisation to ensure cross-infrastructure compatibility (Grüning <i>et al.</i> , 2018) Possibility to execute and share Jupyter notebooks Development repositories available in the Galaxy ToolShed
Scooping	Data and code sharing increases opportunities for collaborations Use pre-print servers to make first claim to a research project "Those who collect data and develop code remain best positioned to undertake future analyses" (pp. 6)	Credit of tools are displayed on the interface Users creating a "Galaxy history" can export a reference list of each tool used, facilitating credit attribution Data can be shared privately through a link while being prepared for publication, or while under embargo.
Lack of time	"Despite the upfront time required, sharing research data and code can ultimately save time for individual researchers and their collaborators, as well as for others who want to reuse it" (pp.7) Begin the research project taking account of future sharing of data and code	More time-consuming in the short term as learning to use a new tool is time-costly but time is saved in the long run as analyses can be re-executed with different parameters, data, or by different users It can help reduce peer review time with possible reproduction of results and easy access to analysis details through the workflow interface
Lack of incentives	"Sharing data and code can increase visibility and recognition of a researcher within the scientific community [...]. It can also help develop open science habits that increase efficiency, and contribute to a better understanding of one's own data and code" (pp.7)	Facilitates sharing and reuse of analytical methods, broader citations of the article associated with the analysis or collaborations could naturally emerge

1078

Acknowledgements

1079

Authors want to thank Sandrine Pavoine for its highly relevant and helpful advices and reviews on both the content and the form of the article.

1080

1081

Authors contribution statement

1082

C. R. drafted the article text, tables, and figures.

1083

C. R. conceptualised the atomisation – generalisation framework with J.-B. M. and Y. L.B. while working on the development of Galaxy workflows.

1085

J.-B. M. and Y. L.B. reviewed and helped rewrite many parts of the draft.

1086

Y. R. and D. P. helped inspire and were invested in the early design of the article.

1087

M. J. and P. S. tested and approved the appliance of the framework.

1088

1089

O. N., M. J., Y. R., M. E., B. B., A. F., H. R. and S. H. highly enhanced the quality of the redaction in both form and content at several stages of the draft.

1090

1091

1092 H. R., S. H., B. B., A. F., and B. G. are involved in the Galaxy-E initiative and
1093 provided many advices on the redaction of the article and/or on the
1094 development of the initiative.

1095 M. E. and G. M. are involved in Antarctic-oriented Galaxy tool and workflow
1096 development coordination.

1097 C. B., R. L., A. M., Y. B., A. A., T. V. and V. C. developed scripts, tools
1098 and/or Galaxy workflows to contribute to the Galaxy-E initiative.

1099 E. A. developed R scripts and apps used to integrate R Shiny apps as
1100 Galaxy interactive tools and initiate "Research Data management Galaxy
1101 tools".

1102 E. M. and C. U. developed the first training materials for Galaxy-E.

1103 E. T. worked on the use of the first Galaxy-E analysis.

1104 M. D., G. L. and R. J. were coordinating the prefiguration of Galaxy-E
1105 through the 65 Millions d'Observateurs project.

1106 Additionnally, all authors reviewed and approved the article draft.

1107 Funding

1108 Funding were provided by the European Union through the Erasmus+
1109 project; the Agence Nationale de la Recherche through the 65 Million
1110 d'Observateurs and the IA-Biodiv projects; the French National Fund for Open
1111 Science through the OpenMetaPaper project; the European commission
1112 through the H2020, the EOSC-Pillar, and the H2020 GAPARS projects; the GO
1113 FAIR initiative through the BiodiFAIRse Implementation Network; the Blue
1114 Nature Alliance; and the Antarctic and Southern Ocean Coalition. Finally,
1115 funding by the French Ministry of Higher Education and Research were
1116 provided for the "Pôle national de données de biodiversité" e-infrastructure.

1117 Conflict of interest disclosure

1118 The authors declare that they comply with the PCI rule of having no
1119 financial conflicts of interest in relation to the content of the article.

1120 References

1121 [Araújo MB, Anderson RP, Barbosa AM, Beale CM, Dormann CF, Early R, Garcia](#)
1122 [RA, Guisan A, Maiorano L, Naimi B, O'Hara RB, Zimmermann NE, Rahbek C](#)
1123 [\(2019\) Standards for distribution models in biodiversity assessments.](#)
1124 [Science Advances, 5, 1-12. <https://doi.org/10.1126/sciadv.aat4858>](#)

1125 Archmiller AA, Johnson AD, Nolan J, Edwards M, Elliott LH, Ferguson JM,
1126 Iannarilli F, Vélez J, Vitense K, Johnson DH, Fieberg J (2020) Computational
1127 Reproducibility in The Wildlife Society's Flagship Journals. *Journal of*
1128 *Wildlife Management*, **84**, 1012-1017. <https://doi.org/10.1002/JWMG.21855>

1129 Batut B, Hiltemann S, Bagnacani A, Baker D, Bhardwaj V, Blank C,
1130 Bretaudeau A, Brillet-Guéguen L, Čech M, Chilton J, Clements D, Doppelt-
1131 Azeroual O, Erxleben A, Freeberg MA, Gladman S, Hoogstrate Y, Hotz HR,
1132 Houwaart T, Jagtap P, Larivière D, Le Corguillé G, Manke T, Mareuil F,
1133 Ramírez F, Ryan D, Sigloch FC, Soranzo N, Wolff J, Videm P, Wolfien M,
1134 Wubuli A, Yusuf D, Taylor J, Backofen R, Nekrutenko A, Grüning B (2018)

1135 Community-Driven Data Analysis Training for Biology. *Cell Systems*, **6**,
1136 752-758. <https://doi.org/10.1016/j.cels.2018.05.012>

1137 Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinel T, Ohl P, Sieb C,
1138 Thiel K, Wiswedel B (2007) KNIME: The Konstanz Information Miner. *Studies*
1139 *in Classification, Data Analysis, and Knowledge Organization*, 319-326.
1140 https://doi.org/10.1007/978-3-540-78246-9_38

1141 Borgman CL (2020) Qu'est-ce que le travail scientifique des données? Big
1142 data, little data, no data. <https://doi.org/10.4000/BOOKS.OEP.14692>

1143 ~~Boyd RJ, August TA, Cooke R, Logie M, Mancini F, Powney GD, Roy DB, Turvey~~
1144 ~~K, Isaac NJB (2023) An operational workflow for producing periodic~~
1145 ~~estimates of species occupancy at national scales. *Biological Reviews*, **98**,~~
1146 ~~1492-1508. <https://doi.org/10.1111/brv.12961>~~

1147 ~~Bray S, Chilton J, Bernt M, Soranzo N, van den Beek M, Batut B, Rasche H,~~
1148 ~~Čech M, Cock PJA, Grüning B, Nekrutenko A (2023) The Planemo toolkit for~~
1149 ~~developing, deploying, and executing scientific data analyses in Galaxy~~
1150 ~~and beyond. *Genome Research*, **33**, 261-268.~~
1151 ~~<https://doi.org/10.1101/gr.276963.122>~~

1152 Carroll S, Garba I, Figueroa-Rodríguez O, Holbrook J, Lovett R, Materechera S,
1153 Parsons M, Raseroka K, Rodriguez-Lonebear D, Rowe R, Sara R, Walker J,
1154 Anderson J, Hudson M (2020) The CARE Principles for Indigenous Data
1155 Governance. *Data Science Journal*, **19**, 43. [https://doi.org/10.5334/dsj-](https://doi.org/10.5334/dsj-2020-043)
1156 [2020-043](https://doi.org/10.5334/dsj-2020-043)

1157 Casajus N. (2023) {rcompendium} {An} {R} package to create a package or
1158 research compendium structure.

1159 Cohen-Boulakia S, Belhajjame K, Collin O, Chopard J, Froidevaux C, Gaignard
1160 A, Hinsén K, Larmande P, Bras Y Le, Lemoine F, Mareuil F, Ménager H,
1161 Pradal C, Blanchet C (2017) Scientific workflows for computational
1162 reproducibility in the life sciences: Status, challenges and opportunities.
1163 *Future Generation Computer Systems*, **75**, 284-298.
1164 <https://doi.org/10.1016/j.future.2017.01.012>

1165 Crusoe MR, Abeln S, Iosup A, Amstutz P, Chilton J, Tijanić N, Ménager H,
1166 Soiland-Reyes S, Goble C (2022) Methods Included: Standardizing
1167 Computational Reuse and Portability with the Common Workflow Language.
1168 *Communications of the ACM*, **65**, 54-63. <https://doi.org/10.1145/3486897>

1169 Culina A, van den Berg I, Evans S, Sánchez-Tójar A (2020) Low availability of
1170 code in ecology: A call for urgent action. *PLOS Biology*, **18**, e3000763.
1171 <https://doi.org/10.1371/JOURNAL.PBIO.3000763>

1172 Di Cosmo R, Zacchiroli S (2017) Software Heritage: Why and How to Preserve
1173 Software Source Code.

1174 Di Tommaso P, Chatzou M, Floden EW, Barja P., Palumbo E, Notredame C
1175 (2017) Nextflow enables reproducible computational workflows. *Nature*
1176 *Biotechnology*, **35**, 316-319. <https://doi.org/10.1038/nbt.3820>

1177 Ellemers N (2021) Science as collaborative knowledge generation. *British*
1178 *Journal of Social Psychology*, **60**, 1-28. <https://doi.org/10.1111/BJSO.12430>

1179 EMBL Australia Bioinformatics Resource (2013) Community Survey Report
1180 [https://www.embl-abr.org.au/news/braembl-community-survey-report-](https://www.embl-abr.org.au/news/braembl-community-survey-report-2013/)
1181 [2013/](https://www.embl-abr.org.au/news/braembl-community-survey-report-2013/)

- 1182 Emery NC, Crispo E, Supp SR, Farrell KJ, Kerkhoff AJ, Bledsoe EK, O'Donnell KL,
1183 McCall AC, Aiello-Lammens ME (2021) Data Science in Undergraduate Life
1184 Science Education: A Need for Instructor Skills Training. *BioScience*, **71**,
1185 1274–1287. <https://doi.org/10.1093/BIOSCI/BIAB107>
- 1186 European Commission, Directorate-General for Research and Innovation
1187 (2018) Cost-benefit analysis for FAIR research data: cost of not having
1188 FAIR research data. *Publications Office*. <https://doi.org/10.2777/02999>
- 1189 Fanelli D (2018) Is science really facing a reproducibility crisis, and do we
1190 need it to? *Proceedings of the National Academy of Sciences of the United*
1191 *States of America*, **115**, 2628–2631.
1192 <https://doi.org/10.1073/pnas.1708272114>
- 1193 Fang FC, Casadevall A (2015) Competitive Science: Is Competition Ruining
1194 Science? *Infection and Immunity*, **83**, 1229–1233.
1195 <https://doi.org/10.1128/IAI.02939-14>
- 1196 Farley SS, Dawson A, Goring SJ, Williams JW (2018) Situating Ecology as a
1197 Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience*,
1198 **68**, 563–576. <https://doi.org/10.1093/BIOSCI/BIY068>
- 1199 Field B, Booth A, Illott I, Gerrish K (2014) *Using the Knowledge to Action*
1200 *Framework in practice: a citation analysis and systematic review*.
1201 *Implementation Science*, **9**, 172. [https://doi.org/10.1186/s13012-014-0172-](https://doi.org/10.1186/s13012-014-0172-2)
1202 [2](https://doi.org/10.1186/s13012-014-0172-2)
- 1203 Goble C, Cohen-Boulakia S, Soiland-Reyes S, Garijo D, Gil Y, Crusoe MR,
1204 Peters K, Schober D (2020) FAIR Computational Workflows. *Data*
1205 *Intelligence*, **2**, 108–121. https://doi.org/10.1162/dint_a_00033
- 1206 Gomes DGE, Pottier P, Crystal-Ornelas R, Hudgins EJ, Foroughirad V, Sánchez-
1207 Reyes LL, Turba R, Martinez PA, Moreau D, Bertram MG, Smout CA, Gaynor
1208 KM (2022) Why don't we share data and code? Perceived barriers and
1209 benefits to public archiving practices. *Proceedings of the Royal Society B*,
1210 **289**, 20221113 <https://doi.org/10.1098/rspb.2022.1113>
- 1211 Gownaris NJ, Vermeir K, Bittner MI, Gunawardena L, Kaur-Ghumaan S,
1212 Lepenies R, Ntsefong GN, Zakari IS (2022) Barriers to Full Participation in
1213 the Open Science Life Cycle among Early Career Researchers. *Data*
1214 *Science Journal*, **21**, 2. <https://doi.org/10.5334/DSJ-2022-002>
- 1215 ~~Green AJ, Figuerola J (2005) Recent advances in the study of long-distance~~
1216 ~~dispersal of aquatic invertebrates via birds. *Diversity and Distributions*, **11**,~~
1217 ~~149–156. <https://doi.org/10.1111/j.1366-9516.2005.00147.x>~~
- 1218 Grüning B, Chilton J, Köster J, Dale R, Soranzo N, van den Beek M, Goecks J,
1219 Backofen R, Nekrutenko A, Taylor J (2018) Practical Computational
1220 Reproducibility in the Life Sciences. *Cell Systems*, **6**, 631–635.
1221 <https://doi.org/10.1016/j.cels.2018.03.014>
- 1222 Hampton SE, Jones MB, Wasser LA, Schildhauer MP, Supp SR, Brun J,
1223 Hernandez RR, Boettiger C, Collins SL, Gross LJ, Fernández DS, Budden A,
1224 White EP, Teal TK, Labou SG, Aukema JE (2017) Skills and Knowledge for
1225 Data-Intensive Environmental Research. *BioScience*, **67**, 546–557.
1226 <https://doi.org/10.1093/BIOSCI/BIX025>
- 1227 Hardisty AR, Bacall F, Beard N, Balcázar-Vargas MP, Balech B, Barcza Z,
1228 Boursat SJ, Giovanni R, Jong Y, Leo F, Dobor L, Donvito G, Fellows D, Guerra
1229 AF, Ferreira N, Fetyukova Y, Fosso B, Giddy J, Goble C, Güntsch A, Haines R,

1230 Ernst VH, Hettling H, Hidy D, Horváth F, Ittzés D, Ittzés P, Jones A,
1231 Kottmann R, Kulawik R, Leidenberger S, Lyytikäinen-Saarenmaa P, Mathew
1232 C, Morrison N, Nenadic A, Hidalgo AN, Obst M, Oostermeijer G, Paymal E,
1233 Pesole G, Pinto S, Poigné A, Fernandez FQ, Santamaria M, Saarenmaa H,
1234 Sipos G, Sylla KH, Tähtinen M, Vicario S, Vos RA, Williams AR, Yilmaz P
1235 (2016) BioVeL: A virtual laboratory for data analysis and modelling in
1236 biodiversity science and ecology. *BMC Ecology*, **16**, 49.
1237 <https://doi.org/10.1186/S12898-016-0103-Y>

1238 Hiltemann S, Rasche H, Gladman S, Hotz HR, Larivière D, Blankenberg D,
1239 Jagtap PD, Wollmann T, Bretaudeau A, Goué N, Griffin TJ, Royaux C, Bras Y
1240 Le, Mehta S, Syme A, Coppens F, Droesbeke B, Soranzo N, Bacon W,
1241 Psomopoulos F, Gallardo-Alba C, Davis J, Föll MC, Fahrner M, Doyle MA,
1242 Serrano-Solano B, Fouilloux AC, van Heusden P, Maier W, Clements D, Heyl
1243 F, Grüning B, Batut B (2023) Galaxy Training: A powerful framework for
1244 teaching! *PLOS Computational Biology*, **19**, e1010752.
1245 <https://doi.org/10.1371/JOURNAL.PCBI.1010752>

1246 Ioannidis JPA (2022) Correction: Why Most Published Research Findings Are
1247 False. *Plos Medicine*, **39**, e1004085.
1248 <https://doi.org/10.1371/JOURNAL.PMED.1004085>

1249 Ivimey-Cook ER, Pick JL, Bairos-Novak K, Culina A, Gould E, Grainger M,
1250 Marshall B, Moreau D, Paquet M, Royauté R, Sanchez-Tojar A, Silva I,
1251 Windecker S (2023) Implementing Code Review in the Scientific Workflow:
1252 Insights from Ecology and Evolutionary Biology. *EcoEvoRxiv*.
1253 <https://doi.org/10.32942/X2CG64>

1254 Jenkins GB, Beckerman AP, Bellard C, Benítez-López A, Ellison AM, Foote CG,
1255 Hufton AL, Lashley MA, Lortie CJ, Ma Z, Moore AJ, Narum SR, Nilsson J,
1256 O'Boyle B, Provete DB, Razgour O, Rieseberg L, Riginos C, Santini L,
1257 Sibbett B, Peres-Neto PR (2023) Reproducibility in ecology and evolution:
1258 Minimum standards for data and code. *Ecology and Evolution*, **13**, e9961.
1259 <https://doi.org/10.1002/ECE3.9961>

1260 [Jetz W, McGeoch MA, Guralnick R, Ferrier S, Beck J, Costello MJ, Fernandez M,](#)
1261 [Geller GN, Keil P, Merow C, Meyer C, Muller-Karger FE, Pereira HM, Regan](#)
1262 [EC, Schmeller DS, Turak E \(2019\) Essential biodiversity variables for](#)
1263 [mapping and monitoring species populations. *Nature Ecology and*](#)
1264 [Evolution, **3**, 539-551. <https://doi.org/10.1038/s41559-019-0826-1>](#)

1265 Keenan M, Cutler P, Marks J, Meylan R, Smith C, Koivisto E (2012) Orienting
1266 international science cooperation to meet global “grand challenges.”
1267 *Science and Public Policy*, **39**, 166-177.
1268 <https://doi.org/10.1093/SCIPOL/SCS019>

1269 Knijn A, Michelacci V, Orsini M, Morabito S (2020) Advanced Research
1270 Infrastructure for Experimentation in genomicS (ARIES): a lustrum of
1271 Galaxy experience. *bioRxiv*. <https://doi.org/10.1101/2020.05.14.095901>

1272 Köster J, Rahmann S (2012) Snakemake—a scalable bioinformatics workflow
1273 engine. *Bioinformatics*, **28**, 2520-2522.
1274 <https://doi.org/10.1093/bioinformatics/bts480>

1275 Lai J, Lortie CJ, Muenchen RA, Yang J, Ma K (2019) Evaluating the popularity of
1276 R in ecology. *Ecosphere*, **10**, e02567. <https://doi.org/10.1002/ECS2.2567>

- 1277 Lamprecht A-L, Garcia L, Kuzak M, Martinez C, Arcila R, Martin Del Pico E,
1278 Dominguez Del Angel V, van de Sandt S, Ison J, Martinez PA, McQuilton P,
1279 Valencia A, Harrow J, Psomopoulos F, Gelpi JL, Chue Hong N, Goble C,
1280 Capella-Gutierrez S (2019) Towards FAIR principles for research software.
1281 *Data Science*, **3**, 37-59. <https://doi.org/10.3233/ds-190026>
- 1282 Larcombe L, Hendricusdottir R, Attwood T, Bacall F, Beard N, Bellis L, Dunn W,
1283 Hancock J, Nenadic A, Orengo C, Overduin B, Sansone S, Thurston M, Viant
1284 M, Winder C, Goble C, Ponting C, Rustici G (2017) ELIXIR-UK role in
1285 bioinformatics training at the national level and across ELIXIR.
1286 *F1000Research*, **6**, 952. <https://doi.org/10.12688/f1000research.11837.1>
- 1287 [Leroy B \(2023\) Choosing presence-only species distribution models. *Journal of*](https://doi.org/10.1111/jbi.14505)
1288 [*Biogeography*, **50**, 247-250. https://doi.org/10.1111/jbi.14505](https://doi.org/10.1111/jbi.14505)
- 1289
- 1290 ~~Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, De Giusti M,~~
1291 ~~L'hours H, Hugo W, Jenkyns R, Khodiyar V, Martone ME, Mokrane M, Navale~~
1292 ~~V, Petters J, Sierman B, Sokolova D V, Stockhause M, Westbrook J (2020)~~
1293 ~~the TRUST Principles for digital repositories. *Scientific Data*, **7**, 144.~~
1294 ~~<https://doi.org/10.1038/s41597-020-0486-7>~~
- 1295 Lortie CJ (2021) The early bird gets the return: The benefits of publishing your
1296 data sooner. *Ecology and Evolution*, **11**, 10736-10740.
1297 <https://doi.org/10.1002/ECE3.7853>
- 1298 McIntire EJB, Chubaty AM, Cumming SG, Andison D, Barros C, Boisvenue C,
1299 Haché S, Luo Y, Micheletti T, Stewart FEC (2022) PERFICT: A Re-imagined
1300 foundation for predictive ecology. *Ecology Letters*, **25**, 1345-1351.
1301 <https://doi.org/10.1111/ELE.13994>
- 1302 Michener WK (2015) Ten Simple Rules for Creating a Good Data Management
1303 Plan. *PLOS Computational Biology*, **11**, e1004525.
1304 <https://doi.org/10.1371/JOURNAL.PCBI.1004525>
- 1305 Michener WK, Jones MB (2012) Ecoinformatics: Supporting ecology as a data-
1306 intensive science. *Trends in Ecology and Evolution*, **27**, 85-93.
1307 <https://doi.org/10.1016/j.tree.2011.11.016>
- 1308 Minocher R, Atmaca S, Bavero C, McElreath R, Beheim B (2021) Estimating
1309 the reproducibility of social learning research published between 1955 and
1310 2018. *Royal Society Open Science*, **8**, 210450.
1311 <https://doi.org/10.1098/RSOS.210450>
- 1312 Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie Du Sert
1313 N, Simonsohn U, Wagenmakers EJ, Ware JJ, Ioannidis JPA (2017) A
1314 manifesto for reproducible science. *Nature Human Behaviour*, **1**, 0021.
1315 <https://doi.org/10.1038/s41562-016-0021>
- 1316 Natural Environment Research Council (2010, 2012) Most Wanted:
1317 Postgraduate Skills Needs in the Environment Sector.
- 1318 Plesser HE (2018) Reproducibility vs. Replicability: A brief history of a
1319 confused terminology. *Frontiers in Neuroinformatics*, **11**, 76.
1320 <https://doi.org/10.3389/FNINF.2017.00076>
- 1321 Powers SM, Hampton SE (2019) Open science, reproducibility, and
1322 transparency in ecology. *Ecological applications*, **29**, e01822.
1323 <https://doi.org/10.1002/eap.1822>

1324 ~~Samota EK, Davey RP (2021) Knowledge and Attitudes Among Life Scientists~~
1325 ~~Toward Reproducibility Within Journal Articles: A Research Survey.~~
1326 ~~Frontiers in Research Metrics and Analytics, 6, 678554.~~
1327 ~~<https://doi.org/10.3389/FRMA.2021.678554>~~

1328 Serrano-Solano B, Fouilloux A, Eguinoa I, Kalaš M, Grüning B, Coppens F
1329 (2022) Galaxy: A Decade of Realising CWFR Concepts. *Data Intelligence*, **4**,
1330 358–371. https://doi.org/10.1162/dint_a_00136

1331 Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández JM,
1332 Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A,
1333 Community R-C, Groth P, Goble C (2022) Packaging research artefacts with
1334 RO-Crate. *Data Science*, **5**, 97–138. <https://doi.org/10.3233/DS-210053>

1335 Strijkers R, Cushing R, Vasyunin D, De Laat C, Belloum ASZ, Meijer R (2011)
1336 Toward executable scientific publications. *Procedia Computer Science*, **4**,
1337 707–715. <https://doi.org/10.1016/j.PROCS.2011.04.074>

1338 The Galaxy Community (2022) The Galaxy platform for accessible,
1339 reproducible and collaborative biomedical analyses: 2022 update. *Nucleic*
1340 *acids research*, **50**, W345–W351. <https://doi.org/10.1093/NAR/GKAC247>

1341 Touchon JC, McCoy MW (2016) The mismatch between current statistical
1342 practice and doctoral training in ecology. *Ecosphere*, **7**, e01394.
1343 <https://doi.org/10.1002/ECS2.1394>

1344 Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A,
1345 Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes
1346 AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R,
1347 Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, t
1348 Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A,
1349 Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA,
1350 Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, Van
1351 Der Lei J, Van Mulligen E, Velterop J, Waagmeester A, Wittenburg P,
1352 Wolstencroft K, Zhao J, Mons B (2016) Comment: The FAIR Guiding
1353 Principles for scientific data management and stewardship. *Scientific Data*,
1354 **3**, 1–9. <https://doi.org/10.1038/sdata.2016.18>

1355 Williams JJ, Teal TK (2017) A vision for collaborative training infrastructure for
1356 bioinformatics. *Annals of the New York Academy of Sciences*, **1387**, 54–60.
1357 <https://doi.org/10.1111/NYAS.13207>

1358 ~~Zurell D, Franklin J, König C, Bouchet PJ, Dormann CF, Elith J, Fandos G, Feng~~
1359 ~~X, Guillera-Arroita G, Guisan A, Lahoz-Monfort JJ, Leitão PJ, Park DS,~~
1360 ~~Peterson AT, Rapacciuolo G, Schmatz DR, Schröder B, Serra-Diaz JM,~~
1361 ~~Thuiller W, Yates KL, Zimmermann NE, Merow C (2020) A standard protocol~~
1362 ~~for reporting species distribution models. *Ecography*, **43**, 1261–1277.~~
1363 ~~<https://doi.org/10.1111/ecog.04960>~~