Title: **How to cook our ecological data: The need for model repositories in ecology**

Corresponding author: Francisco J. Bonet. fjbonet@gmail.com. Tel: +34 958249748.

Summary

1. Most of the information concerning how to create models or use analytic techniques published in the scientific literature is not really available to ecologists. It is stored in scientists' digital or biological memories. We propose that it is time to address the problem of storing, documenting, and executing ecological models and workflows.

2. We are recommending a conceptual framework to design and implement a model repository that will help to address this challenge.

3. We have implemented the conceptual framework in a functional web-based application called ModeleR. This tool is able to document and execute all the models and analytical processes as well as preparing data workflows associated to the Sierra Nevada LTER platform (Spain).

4. We think that model repositories will foster cooperation among scientists, enhancing the creation of relevant knowledge that could be transferred to environmental managers. The overarching idea is to create an international federation of repositories of models with a high degree of connectivity.

Keywords: algorithms, Kepler, metadata, model, model repository, workflow

Introduction: The challenge of storing, documenting and managing algorithms and workflows in ecology

Data analysis, modeling, simulation and other similar methods play a central role in ecology. From statistical models to complex simulation models, the concept of the ecological model has become inherent to the modern ecology. The importance that modeling and data analysis is gaining in ecology is due to at least three main factors.

First, there is a vast amount of primary information available to ecologists to model and analyze. This explosion in data availability (which is georeferenced in some cases), is a result of the enormous effort that ecologists and public agencies have been exerting over the last 3-4 decades in order to gather and share information on the structure and functioning of the Earth's ecosystems. The creation of big data infrastructures that allow ecological data sharing (Whitlock 2011), such as GBIF (Global Biodiversity International Facility), LTER (Long Term Ecological Research) or NEON (National Ecological Observation Network) have greatly contributed to end the era when the lack of primary data restrained the development of ecology as a data-intensive science (Jones *et al.* 2006; Kelling *et al.* 2009; Michener & Jones 2012).

The volume and availability of analytic and modeling methods have undergone an exponential increase in recent decades. Thanks to this evolution, ecologists can choose among dozens of different methodologies to analyze their primary data and design complex models to simulate or analyze the structure or functioning of a given ecosystem.

Lastly, computer science provides a physical (hardware) and logical (software) framework needed to model complex systems and cope with data-intensive computing procedures (Hobbie 2003; Fegraus *et al.* 2005; Plaszczak, Pawel; Wellner, Jr. 2005).

This new data-intensive ecology (called ecoinformatics) was recently described by Michener and Jones (Michener & Jones 2012). They also outlined some of the "remaining challenges" that ecoinformatics faces today. Here, we project our vision of one of these relevant challenges: How should models and algorithms be stored,

documented, and managed in a way that allows their execution and interoperability? Or, in the words of these authors: “little attention has been paid to preserving the algorithms and workflows that scientists use in assuring, analyzing and visualizing data” (Michener & Jones 2012).

Our rationale is that most of the information underlying how to create models or use analytic methods already published in the scientific literature or in technical reports is not readily available to scientists. Most of this knowledge is stored in scientists’ digital or biological memories. Our thesis is that gathering all this knowledge is critical if we truly want the study of Ecology to: a) expand our knowledge of the Earth as a system and our understanding of human impact on that system (Voinov & Cerco 2010); and b) design and implement procedures for sustainable stewardship of natural resources (Chapin *et al.* 2010) in the Anthropocene era (Crutzen 2002). The creation of tools to preserve and manage algorithms and workflows would enhance code sharing and model reuse (Holzworth *et al.* 2010), and would help boost Ecology into taking its place as one of the so-called “big sciences”, whose main features encourage the growth of digital repositories, documentation of data and scientific processes, and creation of technical infrastructures to enhance international collaboration (Borgman *et al.* 2007).

We argue that it is time for ecologists to address the problem of storing, documenting, and executing ecological models. Just as a few decades ago the need for primary-data repositories was obvious; today the creation of repositories of models must be considered the next step in the evolution of Ecology as data-intensive science. A model repository could be defined as an ecoinformatics tool capable of properly managing, documenting, and executing any analytic procedure or workflow created by ecologists.

This challenge is not unique to Ecology, but has also arisen in other areas of science and technology that have followed a similar path, such as Molecular Biology or Earth System Science. These disciplines have evolved from gathering and documenting

primary data to creating complex models, and finally to designing and implementing model repositories that document and execute those models. Preserving workflows, models, and algorithms is, in our opinion, a problem inherent to the "information age" not being exclusive to any given branch of science. This shared history among similar disciplines could provide a valuable set of lessons that would be helpful to all of these fields.

In this work, we propose a conceptual framework to develop a model repository useful for ecologists and environmental managers. This conceptual framework has been built by combining the major advances made by other related scientific areas. We will also illustrate an implementation of this conceptual framework in a real ecological model repository. The result, called ModeleR (Pérez-Pérez, R; Benito, B.M.; Bonet 2012) is the core of an information system that manages the data collected by the global-change monitoring program of Sierra Nevada (Spain) LTER platform. After describing the system envisioned, we will highlight the benefits that this system could offer ecology as a science.

Advances in model repositories

Molecular Biology has advanced in the design and implementation of model repositories. With the immense scientific benefits gained by researchers after the creation of primary-data repositories such as GeneBank, documenting algorithms and models has come to be considered obvious and needed. According to Buckingham (Buckingham 2007), model repositories should allow model documenting (by means of XML schemas or ontologies), must be connected to primary-data repositories, and should be designed using the concept of web service. These concepts have been implemented in several tools (Hunter & Nielsen 2005; Snoep *et al.* 2006). We will highlight Biomodels (Li *et al.* 2010), a repository of peer-reviewed, curated, published,

versionable and parameterizable computational models. To promote the use and growth of Biomodels, some publishers encourage authors to upload their models here after publication. Other initiatives such as myExperiment (Goble *et al.* 2010) enable workflows to be shared among scientists, as in a social network. Among the major contributions made by Molecular Biology to the idea of an ecological model repository are a) the strong connection between primary-data repositories and model repositories and b) the efforts made to promote the use of model repositories in the scientific community.

Earth System Science develops comprehensive and highly integrated models describing the interaction between atmosphere, hydrosphere, lithosphere, biosphere, and heliosphere. This need for integration has prompted the creation of tools similar to model repositories that are able to: a) allow coupling of model execution (Bulatewicz *et al.* 2009; Castronova *et al.* 2012); b) manage model versioning (Thornton *et al.* 2005); c) track computational provenance of models (Frew *et al.* 2008; Dozier & Frew 2009); and d) create models collaboratively thanks to community modeling systems (Voinov *et al.* 2010). Several initiatives have implemented this framework. We will underscore OpenMI (Gregersen *et al.* 2007), which provides a standard interface to describe, document, and execute hydrological models. CSDMS (Peckham *et al.* 2012) (Community Surface Dynamics Modelling System) is able to integrate a wide variety of Earth-surface processes considering different temporal and spatial scales. Major contributions of Earth System Science to the idea of an ecological-model repository have been the advances in model execution and system modularity.

Lastly, Ecology is following this path at a slower pace than the aforementioned disciplines. This could be due to the breadth and diversity of data types in Ecology (Reichman *et al.* 2011). This multiplicity is transferred to the methods that we use to analyze and process those datasets, making it overwhelming to attempt systematization or taxonomy either of data types or of analytic methodologies.

129 Regarding these difficulties, some ecologists have started paying attention to the
130 process of model and algorithm documenting. For example, J. Benz (Benz 1997)
131 wrote a letter in 1997 to the editor of *Ecological Modelling* where he called for a
132 "common model documentation etiquette". Benz proposes a protocol to document
133 ecological models. This protocol has three different levels of documenting: a) general
134 model description (aims, methods, bibliography, authors, links, etc.); b) general
135 information included in the first level plus detailed technical information and some basic
136 information about the mathematics of the model; c) Information included in the second
137 level plus detailed information about the mathematics of the model (algorithm
138 description, parameters, execution rules, etc.). These ideas were partially implemented
139 in an algorithm database called ECOBAS (Hoch *et al.* 1998; Benz & Hoch 2001; Strube
140 *et al.* 2008).

141 **Conceptual framework for an envisioned model repository**

142 Our conceptual framework is explained by describing the basic functions of a model
143 repository (fig. 1): Model documenting is the most important function that a model
144 repository should have. We envision a documenting schema similar to ECOBAS (Benz
145 & Hoch 2001; Strube *et al.* 2008). The "depth" of the documentation process will
146 depend on the knowledge that we have about the model. Two different levels are
147 proposed. First, a basic level implies adding general features such as authorship,
148 rationale, associated bibliography, links to more information, etc. A given model can
149 also be documented in a more thorough way, by adding information about the
150 mathematical operators involved in the model or algorithm. This level of documenting
151 should also include the description of input and output datasets as an inherent part of
152 model's documenting process. Taking into account that input/output datasets are
153 extraordinarily diverse in Ecology (Reichman *et al.* 2011) , the repository should be
154 able to "speak several metadata dialects" (Nogueras-Iso *et al.* 2004) used to document
155 different types of datasets. This could be accomplished thanks to the creation of

156 crosswalks between metadata standards. Thus, the model repository could read
157 metadata catalogs written with different specifications and add these datasets as input
158 data. This feature will allow a high degree of connectivity and modularity between
159 models because an output dataset formed in a given model could be considered an
160 input dataset by another model.

161 Only if the model has reached the most detailed level of documenting, could it be
162 executable by the system. Regarding this function, it is also important to track specific
163 actions by algorithms in input datasets. This feature, called tracking provenance (Frew
164 *et al.* 2008; Dozier & Frew 2009) of the model, allows error tracking and the optimizing
165 of model execution. To facilitate the collaborative process of creating models, the
166 model repository, in our opinion, should include some functions from web 2.0. Blogging
167 or adding comments to models will invite cooperation among modelers who could work
168 together in the processes of model design and implementation.

169 Our proposed approach will be useful only if it can satisfy the needs of scientists as
170 individuals and also research groups to whom they belong. Both the models and the
171 raw data are extremely valuable for ecologists, so that they would prefer to have a sort
172 of local model repository to store and document their models. On the other hand,
173 government agencies or research institutes would be interested in corporate-model
174 repositories able to aggregate models formulated by different researchers. To cope
175 with these scalability needs, we suggest the creation of a federation of local model
176 repositories that are able to communicate to each other using web services. These
177 services would be written using different metadata specifications and would supply
178 information about input/output datasets and algorithm characteristics. This federation of
179 models, based on the statement that information must be stored and curated wherever
180 it is created, is best suited to promote synergy among scientists and also to respect
181 their demands of managing the models that they create. This philosophy has been
182 used to design and implement successful initiatives such as GBIF.

183

184 **ModeleR: a web-based model repository**

185 The conceptual framework that we have described tries to answer the initial question:
186 How should models and algorithms be stored, documented, and managed in a way that
187 allows their execution and interoperability? Besides formulating a conceptual
188 framework, we have also implemented its principal functions. We have created a
189 functional first version of a model repository. This tool, called ModeleR (Pérez-Pérez,
190 R; Benito, B.M.; Bonet 2012) has been built in the context of the Sierra Nevada (Spain)
191 global-change-monitoring program. Sierra Nevada (the highest mountain in the Iberian
192 Peninsula) is a Biosphere reserve, a National Park, and also a LTER platform.

193 ModeleR is able to document any ecological model using different levels of detail
194 shown in the conceptual framework. It also allows the connection of a given model to
195 any dataset documented using EML (Michener 2006). If the model is documented at
196 the most detailed level, the system automatically creates an initial prototype of a Kepler
197 (Altintas *et al.* 2004) workflow. This workflow can be used to execute the model both in
198 a server and in a local computer. However, we can also upload any kind of script that
199 executes the model. We have also implemented the modularity function previously
200 described: model outputs could be used by another model as inputs. To promote the
201 idea of a federation of model repositories, we have created a web service management
202 system that will allow the connection of ModeleR with other repositories. Finally, model
203 creation and documentation can be undertaken collaboratively thanks to a blogging
204 system associated to ModeleR.

205 Currently, we are using ModeleR to document and execute more than 200 models and
206 analytical processes as well as preparing data workflows associated to the Sierra
207 Nevada LTER platform. This tool, which can be accessed via web
208 (http://modeler.obsnev.es/) has become the nucleus of the platform's information

209 system. The global-change-monitoring program under way in Sierra Nevada offers a
210 great breadth of datasets and algorithms that have been used to test both ModeleR as
211 well as the conceptual framework that support it. Fig. 2 presents three examples that
212 are representatives of the amount of models documented in ModeleR.

213

214 **Benefits of model repositories in Ecology**

215 The creation of tools able to document, store and execute ecological models
216 collaboratively (Byrne *et al.* 2010), will contribute to the advance of Ecology as a
217 science in two ways: they will bolster the capacity of creating relevant knowledge and
218 also will improve the capacity that Ecology has to transfer that knowledge to decision
219 makers.

220 Model documenting is a good practice in modeling (Scholten 2008). This task facilitates
221 the process of conceptualization, mathematical description, and real implementation of
222 a given model (Keller & Dungan 1999). A model repository will also help to improve
223 model outreach among different scientists. We believe that model sharing via model
224 repositories will allow ecologists to advance in the reproducibility of ecological analyses
225 (Cassey 2006). Model sharing and code reuse will strengthen the environmental
226 management decisions that could be based upon those models (Scholten 2008). If
227 environmental managers are able to access some ecological models potentially useful
228 in the decision-making process, they will be more likely to try to use them in their area
229 of expertise.

230 We are proposing the creation of tools capable of documenting and executing most
231 types of ecological models or analytical processes. The model repository envisaged will
232 help to extend the paradigm of metadata from datasets to models and algorithms, and
233 will enlarge the toolbox that ecoinformatics is supplying to ecologists.

234 Finally, our approach is also taking into account the scalability needs of individuals,
235 research groups, and institutes. The overarching idea is to create an international
236 federation of repositories of models with a high degree of connectivity. The benefits of
237 this tool are evident and can be summed up with a single word: synergy. The easier it
238 is to share how we "cook" ecological data to get useful knowledge, the stronger the
239 collaboration will be among peers. We propose a web portal where a user could search
240 for hundreds of models and algorithms documented by other ecologists and could read
241 metadata, download workflows, contact other authors to improve models, etc. We have
242 the technology and the knowledge necessary to create such a tool. Now it is up to us.

243

244 **References**

245 Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B. & Mock, S. (2004). *Kepler:*
246 *an extensible system for design and execution of scientific workflows*. IEEE.
247 Retrieved October 11, 2010, from http://www.mendeley.com/research/kepler-an-
248 extensible-system-for-design-and-execution-of-scientific-workflows/

249 Bafna, S., Humphries, J. & Miranker, D.P. (2008). Schema driven assignment and
250 implementation of life science identifiers (LSIDs). *Journal of biomedical*
251 *informatics*, **41**, 730–8. Retrieved June 1, 2012, from
252 http://www.ncbi.nlm.nih.gov/pubmed/18599379

253 Benz, J. (1997). Call for a common model documentation etiquette. *Ecological*
254 *Modelling*, **61**, 141–143. Retrieved April 17, 2012, from http://homepage.ruhr-uni-
255 bochum.de/Michael.Knorrenschild/publ/emletter.ps.gz

256 Benz, J. & Hoch, R. (2001). ECOBAS — modelling and documentation. *Ecological*
257 *Modelling*, **138**, 3–15. Retrieved April 17, 2012, from
258 http://www.sciencedirect.com/science/article/pii/S0304380000003896

Borgman, C.L., Wallis, J.C. & Enyedy, N. (2007). Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, **7**, 17–30. Retrieved March 4, 2012, from http://www.springerlink.com/index/10.1007/s00799-007-0022-9

Buckingham, S. (2007). To build a better model. *Nature Methods*, **4**, 367–373. Retrieved April 9, 2012, from http://www.nature.com/nmeth/journal/v4/n4/box/nmeth0407-367_BX1.html

Bulatewicz, T., Yang, X., Peterson, J.M., Staggenborg, S., Welch, S.M. & Steward, D.R. (2009). Accessible integration of agriculture, groundwater, and economic models using the Open Modeling Interface (OpenMI): methodology and initial results. *Hydrology and Earth System Sciences Discussions*, **6**, 7213–7246. Retrieved from http://www.hydrol-earth-syst-sci-discuss.net/6/7213/2009/

Byrne, J., Heavey, C. & Byrne, P.J. (2010). A review of Web-based simulation and supporting tools. *Simulation Modelling Practice and Theory*, **18**, 253–276. Retrieved March 30, 2012, from http://linkinghub.elsevier.com/retrieve/pii/S1569190X0900149X

Cassey, P. (2006). Reproducibility and repeatability in ecology. *BioScience*, **56**, 958. Retrieved June 1, 2012, from http://www.jstor.org/stable/4488215

Castronova, A.M., Goodall, J.L. & Ercan, M.B. (2012). Integrated modeling within a Hydrologic Information System: An OpenMI based approach. *Environmental Modelling & Software*, 1–11. Retrieved March 15, 2012, from http://linkinghub.elsevier.com/retrieve/pii/S136481521200059X

Chapin, F.S., Carpenter, S.R., Kofinas, G.P., Folke, C., Abel, N., Clark, W.C., Olsson, P., Smith, D.M.S., Walker, B., Young, O.R., Berkes, F., Biggs, R., Grove, J.M.,

283 Naylor, R.L., Pinkerton, E., Steffen, W. & Swanson, F.J. (2010). Ecosystem
284 stewardship: sustainability strategies for a rapidly changing planet. *Trends in*
285 *ecology & evolution*, **25**, 241–9. Retrieved June 14, 2011, from
286 http://www.ncbi.nlm.nih.gov/pubmed/19923035

287 Crutzen, P. (2002). Geology of mankind. *Nature*, **415**, 2002. Retrieved May 3, 2012,
288 from http://academics.eckerd.edu/instructor/carlsopr/Papers/Anthropocene.pdf

289 Dozier, J. & Frew, J. (2009). Computational provenance in hydrologic science: a snow
290 mapping example. *Philosophical transactions. Series A, Mathematical, physical,*
291 *and engineering sciences*, **367**, 1021–33. Retrieved March 14, 2012, from
292 http://www.ncbi.nlm.nih.gov/pubmed/19087938

293 Fegraus, E., Andelman, S. & Jones, M. (2005). Maximizing the value of ecological data
294 with structured metadata: an introduction to ecological metadata language (EML)
295 and principles for metadata creation. *Bulletin of the Ecological Society of America*.
296 Retrieved May 29, 2012, from http://www.esajournals.org/doi/pdf/10.1890/0012-
297 9623(2005)86[158:MTVOED]2.0.CO;2

298 Frew, J., Metzger, D. & Slaughter, P. (2008). Automatic capture and reconstruction of
299 computational provenance. *Concurrency and Computation: Practice and*
300 *Experience*, **20**, 485–496. Retrieved June 3, 2012, from
301 http://onlinelibrary.wiley.com/doi/10.1002/cpe.1247/full

302 Goble, C.A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D.,
303 Borkum, M., Bechhofer, S., Roos, M., Li, P. & De Roure, D. (2010).
304 myExperiment: a repository and social network for the sharing of bioinformatics
305 workflows. *Nucleic acids research*, **38**, W677–82. Retrieved March 9, 2012, from
306 http://nar.oxfordjournals.org/cgi/content/abstract/38/suppl_2/W677

307 Gregersen, J.B., Gijsbers, P.J. a. & Westen, S.J.P. (2007). OpenMI: Open modelling
308 interface. *Journal of Hydroinformatics*, **9**, 175. Retrieved May 10, 2012, from
309 http://www.iwaponline.com/jh/009/jh0090175.htm

310 Hill, C., DeLuca, C. & Suarez, M. (2004). The architecture of the earth system modeling
311 framework. *Computing in Science & Engineering*, 18–28. Retrieved May 17, 2012,
312 from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1255817

313 Hobbie, J.E. (2003). Scientific Accomplishments of the Long Term Ecological Research
314 Program : An Introduction. *BioScience*, **53**, 17–20.

315 Hoch, R., Gabele, T. & Benz, J. (1998). Towards a standard for documentation of
316 mathematical models in ecology. *Ecological Modelling*, **113**, 3–12. Retrieved from
317 http://linkinghub.elsevier.com/retrieve/pii/S0304380098001306

318 Holzworth, D.P., Huth, N.I. & de Voil, P.G. (2010). Simplifying environmental model
319 reuse. *Environmental Modelling & Software*, **25**, 269–275. Retrieved March 9,
320 2012, from http://linkinghub.elsevier.com/retrieve/pii/S1364815208001990

321 Hucka, M., Finney, a., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, a. P.,
322 Bornstein, B.J., Bray, D., Cornish-Bowden, a., Cuellar, a. a., Dronov, S., Gilles,
323 E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr,
324 J.-H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, a., Kummer, U., Le
325 Novere, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D.,
326 Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro,
327 B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner,
328 J. & Wang, J. (2003). The systems biology markup language (SBML): a medium
329 for representation and exchange of biochemical network models. *Bioinformatics*,
330 **19**, 524–531. Retrieved March 8, 2012, from
331 http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btg015

Hunter, P. & Nielsen, P. (2005). A strategy for integrative computational physiology. *Physiology (Bethesda, Md.)*, **20**, 316–25. Retrieved March 15, 2012, from http://www.cfm.brown.edu/crunch/IMAG/Hunter.pdf

Jones, M.B., Schildhauer, M.P., Reichman, O.J. & Bowers, S. (2006). The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 519–544. Retrieved March 2, 2012, from http://www.annualreviews.org/doi/abs/10.1146/annurev.ecolsys.37.091305.110031

Keller, R.M. & Dungan, J.L. (1999). Meta-modeling : a knowledge-based approach to facilitating process model construction and reuse. *Ecological Modelling*, **119**, 89–116.

Kelling, S., Hochachka, W.M., Fink, D., Riedewald, M., Caruana, R., Ballard, G. & Hooker, G. (2009). Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience*, **59**, 613–620. Retrieved March 5, 2012, from http://www.jstor.org/stable/27735945

Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M.I., Snoep, J.L., Hucka, M., Le Novère, N. & Laibe, C. (2010). BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC systems biology*, **4**, 92. Retrieved March 12, 2012, from http://www.biomedcentral.com/1752-0509/4/92

Michener, W.K. (2006). Meta-information concepts for ecological data management. *Ecological Informatics*, **1**, 3–7. Retrieved March 7, 2012, from http://linkinghub.elsevier.com/retrieve/pii/S157495410500004X

356 Michener, W.K. & Jones, M.B. (2012). Ecoinformatics: supporting ecology as a data-
357 intensive science. *Trends in ecology & evolution*, **27**, 85–93. Retrieved March 5,
358 2012, from http://www.ncbi.nlm.nih.gov/pubmed/22240191

359 Miner, R. (2005). The Importance of MathML to Communication. *Notices of the*
360 *American Mathematical Society*, **52**, 532–538.

361 Nogueras-Iso, J., Zarazaga-Soria, F.J., Lacasta, J., Béjar, R. & Muro-Medrano, P.R.
362 (2004). Metadata standard interoperability: application in the geographic
363 information domain. *Computers, Environment and Urban Systems*, **28**, 611–634.
364 Retrieved June 1, 2012, from
365 http://linkinghub.elsevier.com/retrieve/pii/S0198971503001170

366 Oinn, T., Addis, M., Ferris, J., Marvin, D., Carver, T., Pocock, M.R. & Wipat, A. (2004).
367 Taverna: A tool for the composition and enactment of bioinformatics workflows.
368 *Bioinformatics*, **20**, 2004.

369 Peckham, S., Hutton, E. & Norris, B. (2012). A component-based approach to
370 integrated modeling in the geosciences: The design of CSDMS. *Computers &*
371 *Geosciences*. Retrieved May 17, 2012, from
372 http://linkinghub.elsevier.com/retrieve/pii/S0098300412001252

373 Plaszczak, Pawel; Wellner, Jr., R. (2005). *Grid Computing The Savvy Manager's*
374 *Guide*. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA.

375 Pérez-Pérez, R; Benito, B.M.; Bonet, F.J. (2012). ModeleR: An enviromental model
376 repository as knowledge base for experts. *Expert Systems with Applications*, **1**,
377 1829–1841. Retrieved May 3, 2012, from
378 http://onlinelibrary.wiley.com/doi/10.1002/cbdv.200490137/abstract

379 Reichman, O.J., Jones, M.B. & Schildhauer, M.P. (2011). Challenges and opportunities
380 of open data in ecology. *Science (New York, N.Y.)*, **331**, 703–5. Retrieved July 20,
381 2011, from http://www.ncbi.nlm.nih.gov/pubmed/21311007

382 Scholten, H. (2008). Good modelling practice. *13th JISR-IIASA Workshop on*
383 *methodologies and tools*, 15–31. Retrieved April 17, 2012, from
384 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.4951&rep=rep1
385 &type=pdf#page=61

386 Snoep, J.L., Bruggeman, F., Olivier, B.G. & Westerhoff, H.V. (2006). Towards building
387 the silicon cell: a modular approach. *Bio Systems*, **83**, 207–16. Retrieved March 4,
388 2012, from http://www.ncbi.nlm.nih.gov/pubmed/16242236

389 Strube, T., Benz, J., Kardaetz, S. & Brüggemann, R. (2008). ECOBAS — A tool to
390 develop ecosystem models exemplified by the shallow lake model EMMO.
391 *Ecological Informatics*, **3**, 154–169. Retrieved April 17, 2012, from
392 http://linkinghub.elsevier.com/retrieve/pii/S1574954108000058

393 Thornton, P., Cook, R. & Braswell, B. (2005). Archiving numerical models of
394 biogeochemical dynamics. *EOS*, **86**, 6–7. Retrieved April 17, 2012, from
395 http://terraweb.forestry.oregonstate.edu/pubs2/thornton_eos_05.pdf

396 Turuncoglu, U.U., Dalfes, N., Murphy, S. & DeLuca, C. (2012). Toward self-describing
397 and workflow integrated Earth system models: A coupled atmosphere-ocean
398 modeling system application. *Environmental Modelling & Software*, 1–16.
399 Retrieved April 13, 2012, from
400 http://linkinghub.elsevier.com/retrieve/pii/S1364815212000618

401 Voinov, A. & Cerco, C. (2010). Model integration and the role of data. *Environmental*
402 *Modelling & Software*, **25**, 965–969. Retrieved April 12, 2012, from
403 http://linkinghub.elsevier.com/retrieve/pii/S1364815210000435

404 Voinov, A., DeLuca, C., Hood, R. & Peckham, S. (2010). A Community Approach to
405 Earth Systems Modeling. *Eos*, **91**. Retrieved May 17, 2012, from
406 http://192.102.233.13/journals/eo/v091/i013/2010EO13_tabloid.pdf

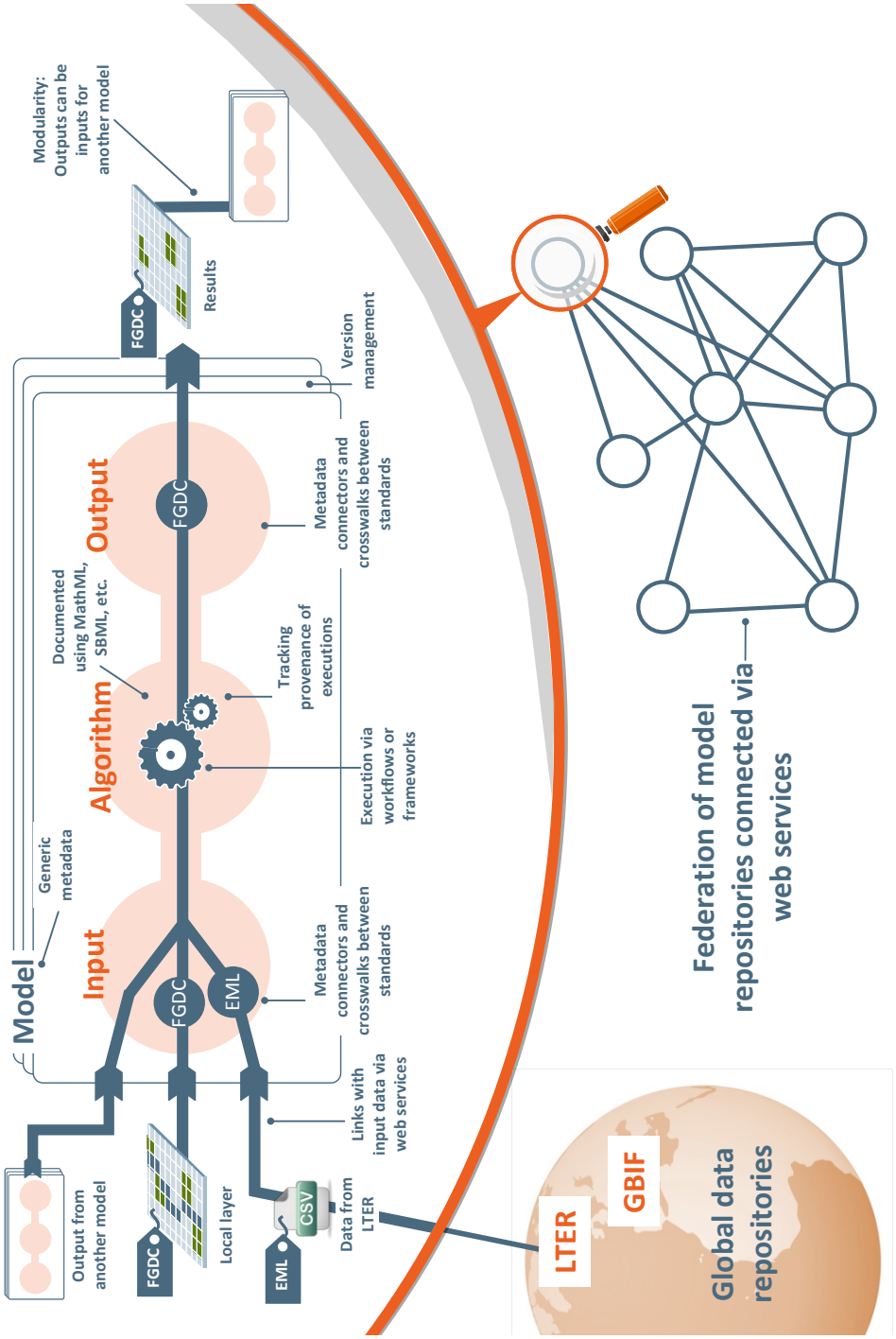
407 Wang, J. (2007). Digital Object Identifiers and Their Use in Libraries. *Serials Review*,
408 **33**, 161–164. Retrieved June 1, 2012, from
409 http://linkinghub.elsevier.com/retrieve/pii/S0098791307000688

410 Whitlock, M.C. (2011). Data archiving in ecology and evolution: best practices. *Trends*
411 *in ecology & evolution*, **26**, 61–5. Retrieved March 7, 2012, from
412 http://www.ncbi.nlm.nih.gov/pubmed/21159406

413

414

415 **Figures**



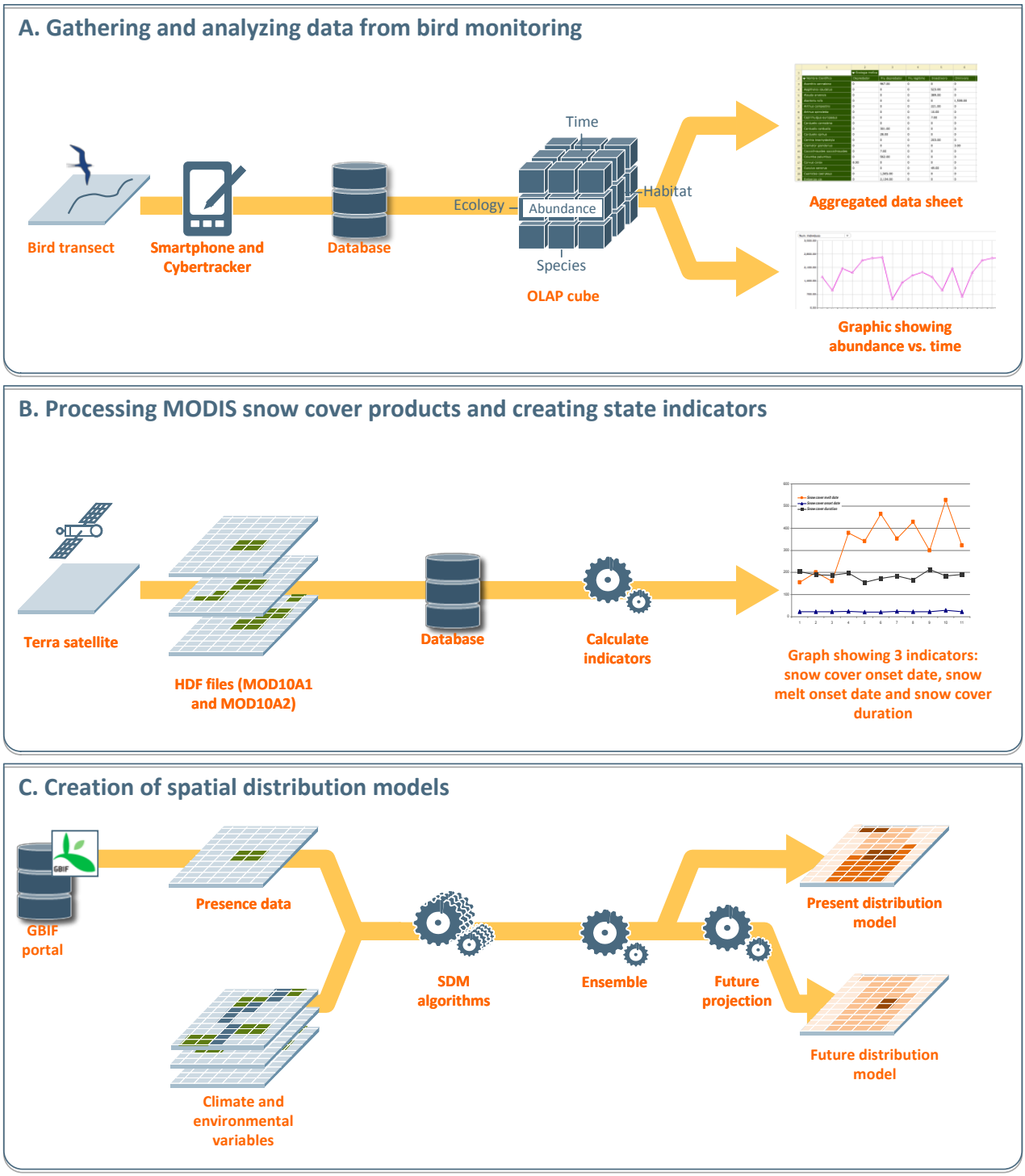
416

417

418 **Figure 1.** Scheme showing the most important functions of the proposed model
419 repository. The envisioned infrastructure should be able to document any ecological
420 model. Different levels of documenting ranging from just adding a name and an unique
421 identifier (using specifications like Life Science Identifiers (Bafna *et al.* 2008), or some
422 others like Digital Object Identifier (Wang 2007)) to a detailed description of
423 mathematical algorithms (by using metadata specifications such us MathML (Miner
424 2005), for equations, or Systems Biology Markup Language (Hucka *et al.* 2003) for
425 biochemical reactions) and input/output datasets (documented via metadata
426 specifications such as Ecological Metadata Language (Michener 2006) and others).
427 Models documented in great detail should be executed by the system. The model
428 repository needs to be able to execute both workflows (like Kepler (Altintas *et al.* 2004)
429 or Taverna (Oinn *et al.* 2004)) and frameworks (like OpenMI (Gregersen *et al.* 2007)). If
430 the model is correctly documented, the system could create a prototype workflow
431 automatically, using metadata from input/output datasets and algorithms. This
432 conceptual framework takes into account the need for a federation of model
433 repositories. Local repositories could connect to some others via web services.

434

435



436 Figure 2. Graphic description of three examples of models documented and executed
437 by ModeleR. The first example (Scheme **A**) shows a set of models and algorithms that
438 allow the gathering of bird-census data. Cybertracker (http://cybertracker.org) yields
439 raw text files via smartphones or PDAs. These files are stored in a PostgreSQL
440 database. ModeleR also documents the process of creating an OLAP cube that allows
441 a multidimensional query system. The outputs of this set of models are sheets and
442 graphics showing the evolution of abundance vs. time, species, habitat type, etc.
443 Scheme **B** shows a workflow that is able to download, process, and create state

444 indicators from MODIS snow products. ModeleR downloads HDF files from NASA's
445 servers. Then ModeleR "explodes" those files and extracts the information for storage
446 in a PostgreSQL database. The last step is the calculation of snow cover indicators
447 using SQL statements. The whole process has been completed combining Kepler
448 workflows, SQL, and ruby on rails. The last example (Scheme **C**) is a workflow to
449 create spatial-distribution models. Presence data are obtained from GBIF portal via
450 web services. Environmental variables are stored in a local computer. Several
451 modeling algorithms are executed using environmental and presence data. Then
452 ensemble forecasting is done to formulate the resulting distribution model. The last
453 step is to project this model into the future by using future climate simulations.

454