Compartiendo datos en Ecología: como añadir (¿más?) valor a tus datos

Andrea Ros Candeira1, Antonio Jesús Pérez-Luque1,

1. Laboratorio de Ecología, Instituto Interuniversitario de Investigación del Sistema Tierra (IISTA-CEAMA), Universidad de Granada, Avda. del Mediterráneo s/n, Granada 18006, España.
2. Grupo de Ecología Terrestre, Departamento de Ecología, Universidad de Granada, Facultad de Ciencias, Campus Fuentenueva s/n, 18071, Granada, España

Autor para correspondencia: A. J. Pérez-Luque [[ajperez@ugr.es](mailto:ajperez@ugr.es)]

# Palabras clave

publicación de datos; repositorios de datos; data papers; metadatos

# Keywords

data publishing; data repository; data papers; metadata

# Introducción

La ciencia puede concebirse como una empresa social cuya finalidad es la acumulación de conocimiento como bien público (Munafò et al., 2017). Uno de los productos de esta gran empresa son, sin duda, los datos. Podría incluso afirmarse que se trata de uno de sus productos más valiosos, teniendo en cuenta que constituyen la base para la generación de información y conocimiento científico (Costello et al., 2013). En la transición hacia una ciencia abierta, la disponibilidad de los datos se perfila como un aspecto inherente a la reproducibilidad (Rodrı́guez-Sanchez et al., 2016), pues contribuye directa y positivamente a que un estudio científico sea más reproducible (Figura 1). En este contexto, la puesta en valor de los datos cobra especial relevancia pues, actualmente, todavía muchos investigadores se encuentran con una baja accesibilidad al recurso. ¿Qué medidas concretas podemos adoptar para maximizar el valor de los datos que generamos y facilitar el acceso a los demás usuarios? En esta nota presentamos algunas ideas para añadir “más” valor a los datos a través de su documentación (metadatado) (Fegraus et al., 2005) y su publicación (Costello et al., 2013) como data papers (Chavan y Penev, 2011).

# La publicación de datos: más allá del material suplementario

En los trabajos científicos, los conjuntos de datos suponen el punto de partida para la realización de análisis y la obtención de resultados, sin embargo, existe la idea equivocada de que un proyecto de investigación finaliza cuando se escribe y publica un artículo científico (Goodman, 2014), dejando en segundo plano el *qué hacer* con los datos originales y su relevancia para la reproducibilidad del estudio. Lo cierto es que la gran mayoría de los estudios científicos no son reproducibles y, en ocasiones, incluso la trazabilidad de los resultados es un misterio para los propios autores (Rodrı́guez-Sanchez et al., 2016). Una deficiente documentación y disponibilidad de los datos en las publicaciones científicas obstaculiza la evaluación crítica de sus resultados (Goodman, 2014). Ante esta situación, incluir los datos como material suplementario de un artículo resulta insuficiente por diversas razones. A modo general, porque la información queda huérfana al no integrarse con el resto de datasets que sustentan el conocimiento científico sobre un tema, dificultando su descubrimiento (Michener, 2015); y a efectos prácticos, porque no puede garantizarse la disponibilidad de los datos a largo plazo, quedando expuestos al devenir de la editorial o revista que los aloja. La baja accesibilidad que afecta a los datos incluidos como material suplementario se debe a diversas causas: la diversidad de formatos utilizada por los autores para enviar sus datos; la falta de estándares que armonicen los conjuntos de datos; la inexistencia de herramientas de búsqueda que imposibilitan su localización, etc. (Santos et al., 2005). Resulta evidente que incluir los datos de este modo no sustituye, de ninguna manera, su depósito en un repositorio público.

En el caso concreto de la disciplina ecológica, un incremento en la accesibilidad a los datos requiere la superación de dos retos: el tecnológico y el social (Reichman et al., 2011). El primer aspecto guarda relación con la complejidad inherente a la información ecológica mientras que, el segundo, refleja la necesidad de superar ciertas barreras sociales todavía presentes en la ciencia a la hora de compartir datos. Tal como señala Michener (2015), las revistas y las editoriales pueden actuar como impulsores de este cambio social adoptando políticas de datos que exijan a los autores la publicación de sus datos en abierto e, incluso, fomentando la creación de revistas de datos (Tabla 2). Una iniciativa importante tuvo lugar hace 9 años, cuando varias revistas clave en el campo de la evolución y la ecología adoptaron una nueva política de datos llamada *Joint Data Archiving Policy* (Whitlock, 2011), la cual introducía como requisito indispensable para la publicación del artículo la accesibilidad a los datos que sustentaban la investigación. El panorama actual es variado, pero los cambios en las políticas de datos de las revistas muestran una clara tendencia general hacia una mayor disponibilidad de datos (Stodden, 2013).

Para hacernos una idea, en una revisión de las políticas de datos de las revistas de mayor impacto en el campo de la Ecología (Tabla 1), hemos observado la existencia de un apoyo explícito (ej.: *Ecology Letters*) e implícito (ej.: *Methods in Ecology and Evolution* y *Ecological Monographs*) de algunas revistas a la reproducibilidad, poniendo énfasis en la necesidad de hacer públicos los datos para su reutilización, y exigiendo su depósito en un repositorio público como condición de publicación. Algunas, de hecho, son críticas con el uso del material suplementario como alternativa. Sin embargo, llama la atención el diferente grado de expectativa a la hora de compartir datos. Por ejemplo, en las ciencias ómicas, se espera que los investigadores depositen sus datos, como las secuencias de ADN; mientras que es opcional para otro tipo de datos (e.j. *The ISME Journal*). Este “trato diferente” evidencia que algunas disciplinas ya han superado los retos tecnológicos y sociales que en ecología aún son tarea pendiente. Un ejemplo vivo de este cambio de paradigma se manifiesta en repositorios como GenBank (Reichman et al., 2011). La proliferación de éste y muchos otros repositorios ha jugado un papel crucial en la disponibilidad de datos. La variedad de repositorios a los que tenemos acceso los investigadores es muy alta, desde repositorios más generalistas que almacenan datos de diversas ramas científicas, como por ejemplo Dryad (<https://datadryad.org>); a otros más específicos, como GBIF (Global Biodiversity Information Facility, <https://www.gbif.org>) para datos de biodiversidad o el ITRDB (International Tree-Ring Data Bank) (<https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring>) para datos dendrocronológicos. Para ayudarnos a identificar el repositorio más apropiado existen herramientas como OpenDOAR (<http://v2.sherpa.ac.uk/opendoar/>) o RepositoryFinder (<https://repositoryfinder.datacite.org/>), que realiza búsquedas en el registro internacional de repositorios de datos de investigación re3data (<https://www.re3data.org/>). Asimismo, existen listados de repositorios categorizados por disciplinas (<https://www.nature.com/sdata/policies/repositories>) y por tipo de datos (e.j. Thessen y Patterson, 2011).

# Los metadatos: háblame de tus datos y…

Para promover la reutilización de los datos y la reproducibilidad, resulta insuficiente el simple depósito de los datos en un repositorio. Es necesario dotar a los datos de una mínima información adicional (metadatos) (F. Valladares, 2006) ya que, además de una buena defensa contra el olvido y el tiempo que pasa desde su colecta, documentar los datos ayuda a la interpretación correcta de los mismos y hace viable su reutilización por otros usuarios (Fegraus et al., 2005). En este sentido, los principios FAIR (*Findable*, *Accessible*, *Interoperable*, and *Reusable*) (Wilkinson et al., 2016) promueven que los datos, además de estar accesibles, estén bien documentados para poder reutilizarse.

Para facilitar el intercambio, la integración y síntesis de los datos se han desarrollado diversos estándares de metadatos, en función del propósito y la disciplina científica. En una revisión para la disciplina ecológica, Alves et al. (2018) encontraron que los más utilizados son: Ecological Metadata Language (EML) (<http://knb.ecoinformatics.org/software/eml/>); Darwin Core (Wieczorek et al., 2012); y la Directiva Europea INSPIRE para datos espaciales (<https://inspire.ec.europa.eu/>).

Para la creación de metadatos existe un amplio abanico de herramientas. La red europea de seguimiento a largo plazo LTER (Long-Term Ecological Research) ha creado una herramienta llamada DEIMS-SDR (Dynamic Ecological Information Management System - Site and dataset registry) (<https://deims.org/>) (Wohner et al., 2019), que permite documentar conjuntos de datos ecológicos y proporcionarlos en diferentes estándares (EML, Biological Data Profile, INSPIRE). En R existen varios paquetes que nos permiten trabajar con el estándar EML, entre ellos destacamos los paquetes EML (<https://ropensci.github.io/EML/>) y emld (<https://cboettig.github.io/emld/>) para la creación de metadatos, y el paquete EMLAssemblyLine (<https://github.com/EDIorg/EMLassemblyline>) que incorpora además un flujo de trabajo. Existen asimismo, otros que facilitan la visualización de los metadatos creado páginas web sencillas para ello, como el paquete emldown (<https://github.com/ropenscilabs/emldown>).

Algunos investigadores ya están incorporando esta aproximación de documentación de datos y depósito de los mismos en repositorios internacionales siguiendo el estándar EML. Por ejemplo Parmenter et al. (2018), en un trabajo sobre factores ambientales y endógenos que afectan a la vecería en bosques de Nuevo México, realizó una documentación de los datos mediante el estándar EML y su posterior integración dentro de repositorio de la Red LTER.

# El data paper: conjugando datos y metadatos en una publicación científica

Ir un paso más allá de publicar los datos junto con los metadatos en un repositorio puede ser escribir un data paper sobre ese conjunto de datos. El data paper es una publicación científica que documenta detalladamente uno o varios conjuntos de datos accesibles, describiendo el contexto en el que fueron generados y su contenido. Respecto a su estructura, no sigue el esquema típico de un artículo científico, pues su propósito no es exponer una investigación, sino describir conjuntos de datos de forma entendible y estructurada. El data paper es, por tanto, un producto que pone en valor el recurso de datos que documenta, dándole visibilidad a su contenido y reconocimiento académico a sus proveedores (Chavan y Penev, 2011). Costello et al. (2013) resalta esta modalidad de publicación como una garantía de accesibilidad y calidad de los datos y metadatos, ya que implica una revisión por pares del producto. Siguiendo esta línea, en los últimos años ha aumentado el número de revistas dedicadas exclusivamente a la publicación de data papers (data journals) y, con mayor frecuencia, revistas de larga trayectoria incorporan el data paper como modalidad de publicación (Tabla 2).

# Aproximación “tradicional” *vs.* Aproximación reproducible. Un ejemplo: diversidad en los bosques de Sierra Nevada.

Sierra Nevada es una región montañosa situada en el sur de Europa, considerada un hotspot de diversidad vegetal en la cuenca Mediterránea. Alberga 2353 taxones de flora vascular, que constituye el 33.2 % de la flora de España (incluyendo las islas) (Lorite, 2016). En esta región, que cuenta con una larga tradición de seguimiento de los procesos ecológicos a largo plazo (Pérez-Luque et al., 2016), en el año 2004 se realizó un exhaustivo inventario forestal en 600 parcelas, distribuidas a lo largo de gradientes ecológicos y altitudinales. Este inventario, conocido como Sinfonevada (Pérez-Luque et al., 2014), fue utilizado por diversos investigadores para avanzar en el campo de la ecología de las repoblaciones (Apéndice 1), identificando los factores ecológicos clave que afectan a la regeneración natural bajo repoblaciones de coníferas, como por ejemplo: la proximidad a la fuente semillera, las características intrínsecas de la plantación o la importancia del uso del suelo previo a la repoblación (Gómez-Aparicio et al., 2009; González-Moreno et al., 2011; Navarro-González et al., 2013).

Inicialmente, este conjunto de datos se utilizó siguiendo una aproximación que hemos denominado “tradicional” respecto al manejo de los datos. Esto es, se llevó a cabo un análisis de los datos del inventario forestal (parte de ellos o en su totalidad) para testar algunas hipótesis y publicar los resultados en alguna revista científica, añadiendo los datos como material suplementario. Esta aproximación en el tratamiento del conjunto de datos se limitó al depósito como material suplementario; y aunque se mejoró el acceso potencial a dichos datos (179 citas de los tres trabajos que ha utilizado directamente parte de este conjunto de datos) (Figura 3a) su utilización estaba limitada por la escasa documentación sobre dicho inventario, que aparecía a lo sumo dispersa y no estructurada.

Para mejorar el acceso a dicho conjunto de datos, se llevó a cabo una aproximación que complementa el uso de los datos, añadiendo más valor a los mismos, y aumentando la posibilidad de reutilización del conjunto original de datos, para otros fines diferentes del propósito inicial. Esta aproximación “reproducible” consistió en la normalización del inventario forestal y su posterior integración en repositorios de datos (Figura 3b). Para ello se realizó una documentación del conjunto de datos que permite conocer todos los detalles de los datos (desde su génesis hasta su “cocinado”), y posteriormente se integró en un repositorio local. Además, para poder integrarlo en iniciativas internacionales y que otros investigadores pudieran conocer el conjunto de datos, se utilizó un estándar de documentación (EML y Darwin Core). De tal forma que el conjunto de datos inicial, una vez normalizado y documentado de acuerdo a un estándar de metadatos, está disponible en repositorios internacionales, y puede ser utilizado para fines diferentes. En nuestro ejemplo, el conjunto de datos Sinfonevada integrado en GBIF, ha sido citado 36 veces, y lo más interesante es que se ha descargado (total o parcialmente) más de 11267 veces (Figura 3d), aumentando en varios órdenes de magnitud el uso inicial de dicho conjunto de datos. Además de la inclusión del conjunto de datos en un repositorio institucional, se consideró complementario la opción de una descripción detallada del conjunto de datos, para lo cual se realizó un Data Paper, que tras un proceso de revisión se publicó en una revista indexada (Figura 3c).

A nuestro entender, esta forma de compartir datos (ecológicos o de otras disciplinas), conlleva una serie de beneficios, entre los que destacamos:

* Aumento de la longevidad de los datos. Al documentar el conjunto de datos e integrarlos en repositorios institucionales evitamos el peligro de que se pierdan los datos en el cajón de los investigadores o que se pierdan por rotura de discos de almacenamiento.
* Control de calidad de los datos. Durante la documentación de los datos se realizan varios procesos de control de calidad de los datos.
* Aumento de la accesibilidad a los datos, tanto a nivel local como internacional.
* Incremento de la reutilización de los datos con propósitos diferentes a los originales.

Como vemos existen muchos argumentos a favor de archivar y documentar los datos, y no son pocos los ejemplos donde los datos se reutilizan tras varias décadas para unos propósitos diferentes a los que motivaron su colecta (Specht et al., 2018). Sin embargo, no podemos olvidarnos de los costes, tanto en tiempo como en dinero, asociado a la puesta en valor de los datos pasados y actuales (Costello et al., 2013; Michener, 2015, @Goodhill2014; Specht et al., 2018), siendo necesario además un reconocimiento a este trabajo.

# Conclusiones

Al igual que las especies, existen datos que se encuentran *“en peligro de extinción”* y es importante dedicar esfuerzos a su preservación, mediante su documentación y depósito en repositorios oficiales (Specht et al., 2018). ¿Cómo hacemos más sencillo este camino? El reto no sólo está en manos de las revistas y las editoriales, también los investigadores podemos contribuir a ello documentando los datos, publicándolos y apostando por productos como los data papers. En los últimos años han surgido diferentes recomendaciones y buenas prácticas que aportan consejos sobre la gestión de los datos y cómo compartirlos, muchos de ellas centrados en la disciplina ecológica (Kervin et al., 2013; Goodman, 2014; Michener, 2015). Estos hábitos, saludables en términos de reproducibilidad, posibilitan la reutilización de los datos dando lugar a nuevos estudios capaces de lograr tanto impacto como las investigaciones originales o incluso reutilizándolos con un propósito insospechado en el momento de su colecta, nutriendo así la ciencia del futuro (Editorial, 2018; Specht et al., 2018). Los esfuerzos por hacer de los datos y sus metadatos productos publicables y citables están encaminando una puesta en valor de la información que repercute en una serie de beneficios no sólo para el investigador, sino para la comunidad científica en general.

Finalmente, tras revisar normas de publicación de la Revista Ecosistemas, no hemos encontrado ninguna recomendación y/o política de datos. En este sentido, consideramos que sería muy positivo que la revista apoye de forma explícita la documentación y puesta en valor de los datos, así como su inclusión en repositorios oficiales, y además sería interesante la creación de una sección de DataPapers en la revista.

# Agradecimientos

Los autores agradecen al grupo de Ecoinformática por la revisión de la Nota, y a los proyectos ECOPOTENTIAL (EU Horizon 2020 grant agreement No 641762): *Improving future ecosystem benefits through earth observations*; y LIFE-ADAPTAMED (LIFE14 CCA/ES/000612): *Protección de servicios ecosistémicos clave amenazadas por el cambio climático mediante gestión adaptativa de socioecosistemas mediterráneos*. Este trabajo ha sido desarrollado bajo el *Convenio de colaboración entre la Consejería de Medio Ambiente y Ordenación del Territorio de la Junta de Andalucía y la Universidad de Granada para el desarrollo de actividades vinculadas al Observatorio de Cambio Global de Sierra Nevada*.

# Referencias

Alves, C., Castro, J., Ribeiro, C., Honrado, J., Lomba, Â. 2018. Research data management in the field of Ecology: an overview. *International Conference on Dublin Core and Metadata Applications* 0: 87-94.

Chavan, V., Penev, L. 2011. The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics* 12 Suppl 15: S2-S2.

Costello, M.J., Michener, W.K., Gahegan, M., Zhang, Z.-Q., Bourne, P.E. 2013. Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution* 28: 454-461.

Editorial. 2018. Data sharing and the future of science. *Nature Communications* 9: 2817.

Fegraus, E.H., Andelman, S., Jones, M.B., Schildhauer, M. 2005. Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. *The Bulletin of the Ecological Society of America* 86: 158-168.

F. Valladares, B.A. y. 2006. Bases de datos y metadatos en ecología: compartir para investigar en cambio global. *Revista Ecosistemas* 15:.

González-Moreno, P., Quero, J., Poorter, L., Bonet, F., Zamora, R. 2011. Is spatial structure the key to promote plant diversity in Mediterranean forest plantations? *Basic and Applied Ecology* 12: 251-259.

Goodhill, G.J. 2014. Practical costs of data sharing. *Nature* 509: 33-33.

Goodman, A.A.B., Alyssa AND Pepe. 2014. Ten Simple Rules for the Care and Feeding of Scientific Data. *PLOS Computational Biology* 10: 1-5.

Gómez-Aparicio, L., Zavala, M.A., Bonet, F.J., Zamora, R. 2009. Are pine plantations valid tools for restoring Mediterranean forests? An assessment along abiotic and biotic gradients. *Ecological Applications* 19: 2124-2141.

Kervin, K., Michener, W., Cook, R. 2013. Common Errors in Ecological Data Sharing. *Journal of eScience Librarianship* 2: e1024.

Lorite, J. 2016. An updated checklist of the vascular flora of Sierra Nevada (SE Spain). *Phytotaxa* 261: 1-57.

Michener, W.K. 2015. Ecological data sharing. *Ecological Informatics* 29: 33-44.

Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U. et al. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1: 0021 EP.

Navarro-González, I., Pérez-Luque, A.J., Bonet, F.J., Zamora, R. 2013. The weight of the past: land-use legacies and recolonization of pine plantations by oak trees. *Ecological Applications* 23: 1267-1276.

Parmenter, R.R., Zlotin, R.I., Moore, D.I., Myers, O.B. 2018. Environmental and endogenous drivers of tree mast production and synchrony in piñon–juniper–oak woodlands of New Mexico. *Ecosphere* 9: e02360.

Pérez-Luque, A.J., Bonet, F.J., Pérez-Pérez, R., Aspizua, R., Lorite, J., Zamora, R. 2014. Sinfonevada: Dataset of floristic diversity in Sierra Nevada forests (SE Spain). *PhytoKeys* 35: 1-15.

Pérez-Luque, A.J., Bonet, F.J., Zamora, R., Barea-Azcón, J.M., Aspizua, R., Sánchez-Gutiérrez, F.J. 2016. Impacts of global change in the Sierra Nevada LTER-site. *Revista Ecosistemas* 25: 65-71.

Reichman, O.J., Jones, M.B., Schildhauer, M.P. 2011. Challenges and Opportunities of Open Data in Ecology. *Science* 331: 703-705.

Rodrı́guez-Sanchez, F., Pérez-Luque, A.J., Bartomeus, I., Varela, S. 2016. Ciencia reproducible: qué, por qué, cómo? *Revista Ecosistemas* 25:.

Santos, C., Blake, J., States, D.J. 2005. Supplementary data need to be kept in public repositories. *Nature* 438: 738 EP.

Specht, A., Bolton, M.P., Kingsford, B., Specht, R.L., Belbin, L. 2018. A story of data won, data lost and data re-found: the realities of ecological data preservation. *Biodiversity Data Journal* 6: e28073.

Stodden, P.A.M., Victoria AND Guo. 2013. Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. *PLOS ONE* 8: 1-8.

Thessen, A., Patterson, D. 2011. Data issues in the life sciences. *ZooKeys* 150: 15-51.

Whitlock, M.C. 2011. Data archiving in ecology and evolution: best practices. *Trends in Ecology & Evolution* 26: 61-65.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D. 2012. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard Sarkar, I. N. (ed.), *PLoS ONE* 7: e29715.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018.

Wohner, C., Peterseil, J., Poursanidis, D., Kliment, T., Wilson, M., Mirtl, M., Chrysoulakis, N. 2019. DEIMS-SDR – A web portal to document research sites and their associated data. *Ecological Informatics* 51: 15-24.

###### TABLA 1

**Tabla 1**. Requerimientos en materia de datos de las revistas con mayor impacto en ecología. Búsqueda de revistas en Web of Science (2019/01/23) y revisión de política de datos para aquellas cuya publicación principal es el artículo de investigación.

**Table 1.** Data sharing requirements of the journals with the greatest impact on Ecology. A search of journals by rank using Web of Science (2019/01/23) and data policy review for those whose main publication is the research article.

###### TABLA 2

**Tabla 2.** Ejemplo de revistas que aceptan artículos de datos como modalidad de publicación.

**Table 2.** Example of journals that accept data papers as publication modality.

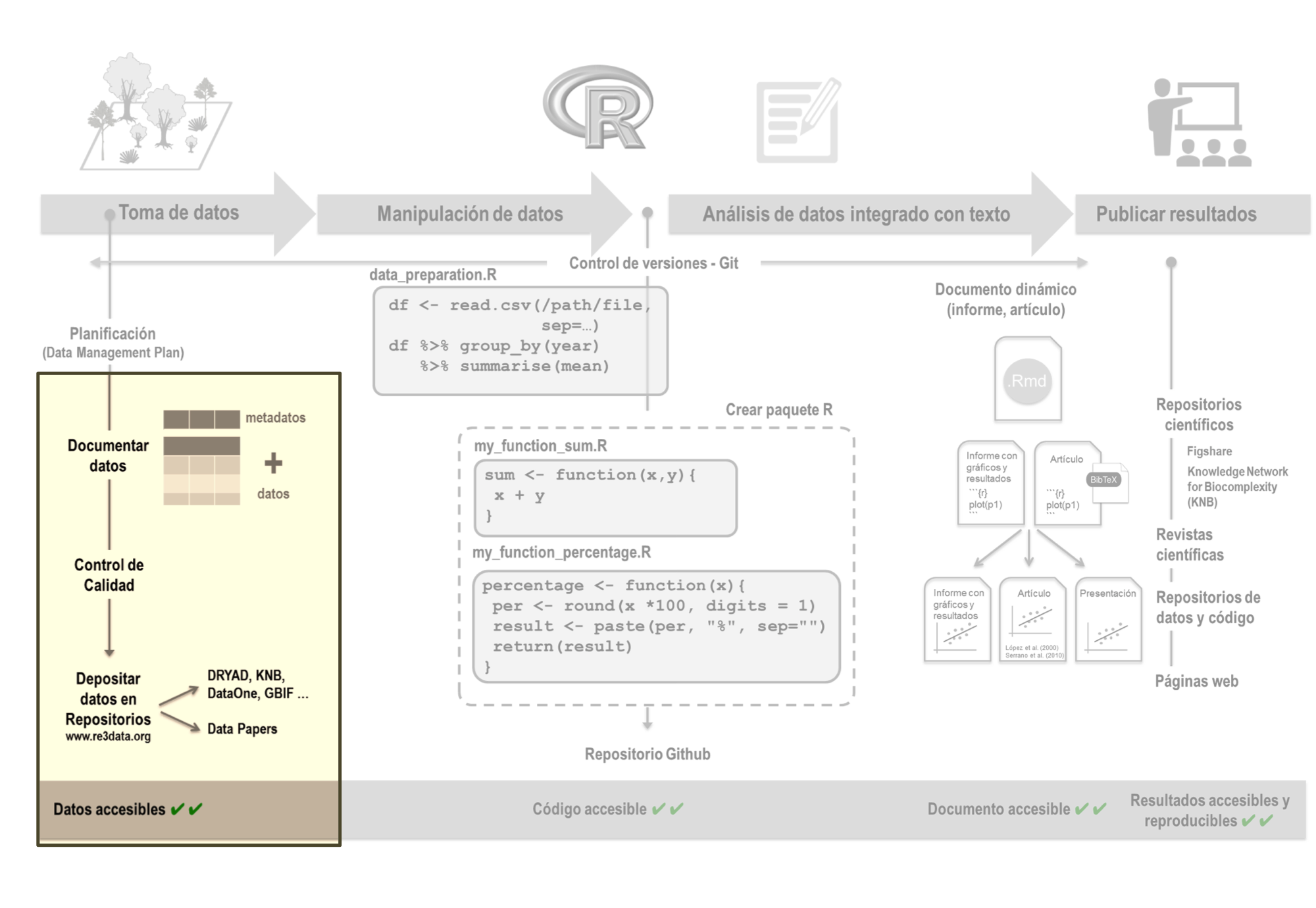
###### PIES DE FIGURA

**Figura 1**. Esta nota pone el foco en las acciones que suponen el punto de partida para un flujo de trabajo reproducible. Adaptado de Rodrı́guez-Sanchez et al. (2016).

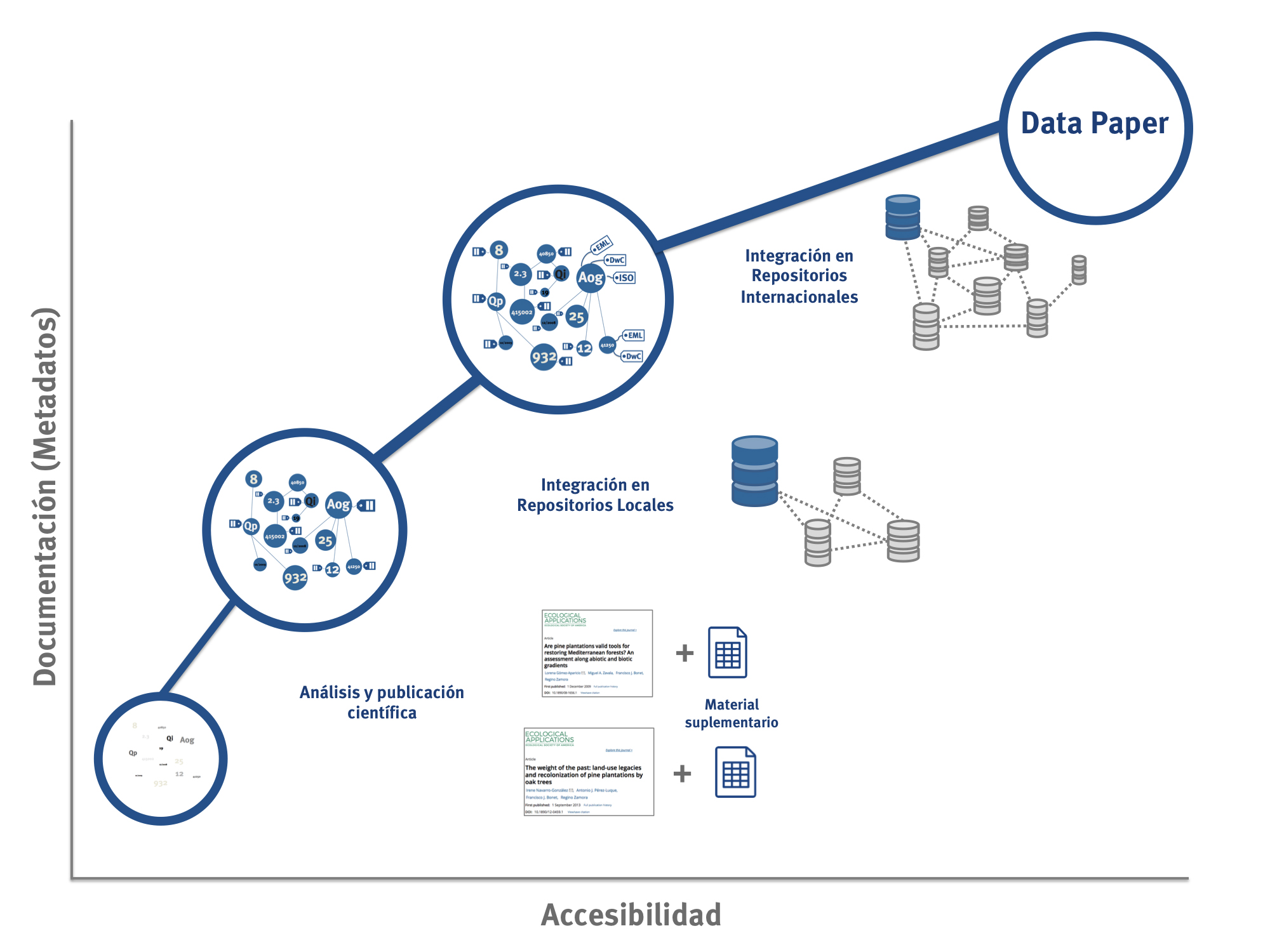
###### FIGURE LEGENDS

**Figure 1**. This publication focuses on actions that are the baseline for a reproducible workflow. Adapted from Rodrı́guez-Sanchez et al. (2016).

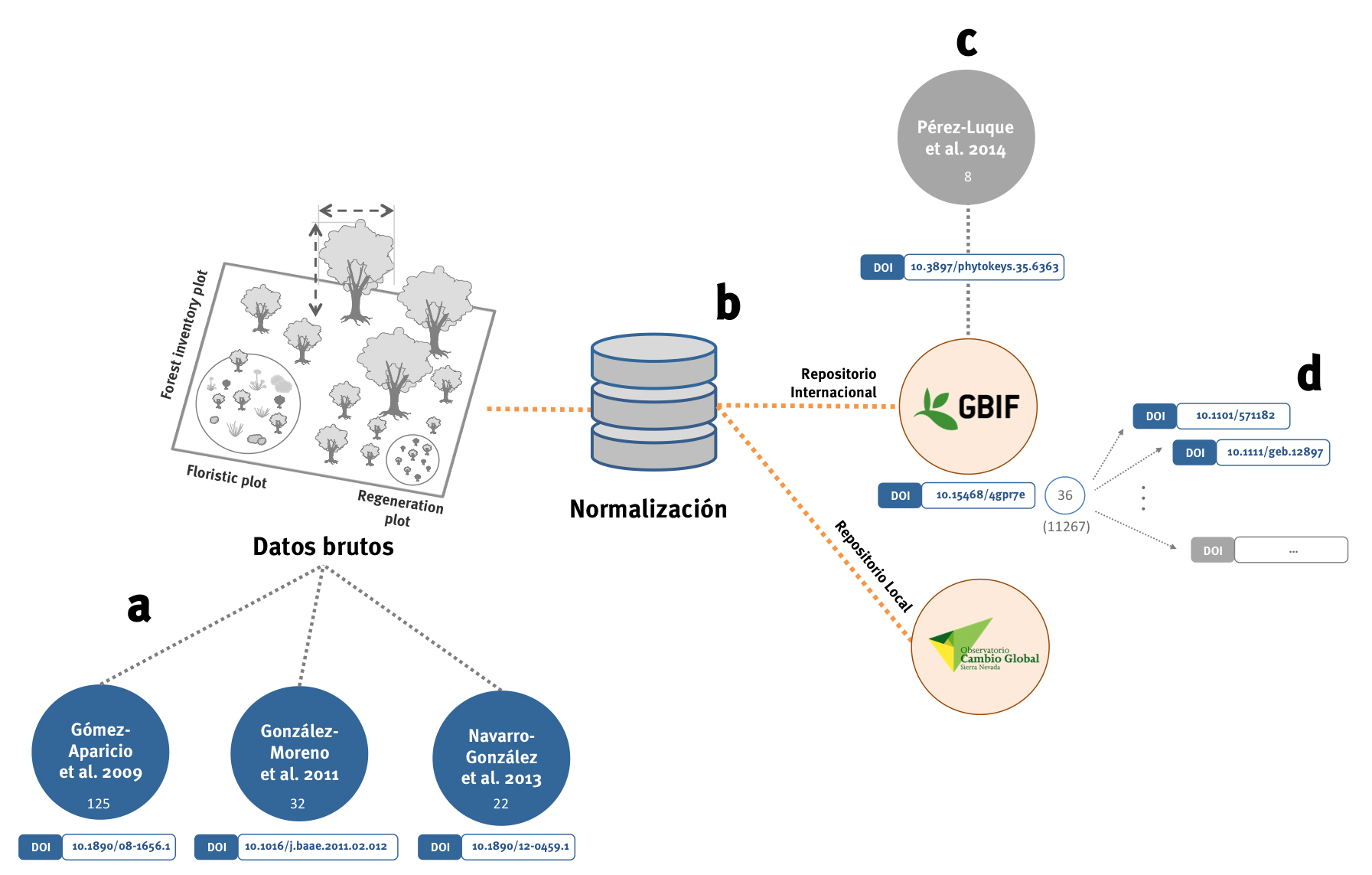
###### FIGURA 1



###### FIGURA 2



###### FIGURA 3



**Apéndice 1. Publicaciones que han utilizado el Inventario Forestal Sinfonevada antes de su integración en un repositorio.**

**Appendix 1. Publications that have used the Sinfonevada Forest Inventory before its integration into a repository.**