Universidad de Navarra

Facultad de Ciencias
Departamento de Zoología y Ecología

Metaanálisis de los datos de biodiversidad suministrados a través de gbif.es

Resumen

Los registros de datos que diversos proveedores hacen públicos a través de gbif.es pueden considerarse una representación del conocimiento de la biodiversidad generado en España. Analizando con técnicas variadas o novedosas de exploración qué pautas se observan en esta masa de datos, es posible detectar qué fortalezas y carencias hay en este conocimiento, y cómo está evolucionando. Hemos encontrado tres grupos de pautas de interés: cobertura geográfica, extensión temporal, y preferencia taxonómica. La evolución de la calidad datos mejora en general con el tiempo, aunque también detectamos algún componente que ha empeorado. Las técnicas empleadas aquí pueden servir para diseñar mecanismos de monitorización de la calidad de los datos servidos y para calibrar su utilidad para la investigación y gestión, así como para diseñar un servicio público de consulta accesible desde gbif.es.

Introducción

En octubre de 2009 GBIF.ES planteó examinar la masa de datos que son servidos desde la unidad de coordinación (nodo nacional) de GBIF en España (gbif.es), y que son la parte más representativa de la contribución española al esfuerzo global representado por gbif.org.

Interesaba, sobre todo, disponer de una visión de conjunto de los datos entregados; poner a punto técnicas analíticas, y estimar de su calidad especialmente en cuanto a su potencial uso en investigación y gestión.

Una vez realizado este examen preliminar, las técnicas desarrolladas pueden aplicarse a análisis de mayor volumen, como el conjunto de datos servidos por otros grupos de proveedores (por ejemplo, proveedores españoles). En UNAV estamos actualmente desarrollando estas técnicas y aplicándolas en el conjunto de datos de GBIF.org.

Los registros primarios de biodiversidad que sirve GBIF pertenecen a numerosos proveedores, los cuales han generado estos datos primarios de ocurrencias (generalmente, ejemplares capturados o depositados en colecciones) y observaciones de campo. Aunque una parte muy importante de este tipo de datos no se ha digitalizado aún o se encuentra disponible sólo a través de medios clásicos de información (publicaciones o informes), cabe suponer que lo que se está haciendo disponible para consulta inmediata a través de las redes de datos podría representar, como si fuera un muestreo, al conjunto del conocimiento sobre la biodiversidad generado en un ámbito.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España

Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

Bajo esta hipótesis, al analizar las estructuras emergentes en los datos accesibles a través de gbif.es—que son mayoritariamente los que suministran los proveedores de datos españoles que, a su vez, representan a las entidades que investigan la biodiversidad—, las pautas que pudieran encontrarse podrían representar las pautas sobre el conocimiento de la biodiversidad, sea ésta de España o aquél generado en España.

Como todo muestreo, la fiabilidad de las conclusiones que se derivan de él dependerá de lo representativo que sea de aquello que se muestrea. Gbif.es recoge datos de diversas fuentes, pero sólo de las que han decidido hacer públicos sus datos primarios a través de este medio y posen tales datos en un formato adecuado para ello. Estas dos condiciones no pueden darse por garantizadas para una muestra aleatoria de todas las entidades que generan datos de biodiversidad, por diversos motivos. Sin embargo, parece evidente que sí puede considerarse una muestra representativa de las entidades que *publican datos digitalizados* de biodiversidad, que son, en general, los que resultan utilizables de forma más inmediata por científicos, técnicos y gestores, al estar (por definición) publicados y existir en formato digital, fácilmente asequible hoy en día por medios informáticos a través de las redes de información.

Este estudio examina los datos servidos por gbif.es atendiendo a varios factores que afectan a su calidad, y por tanto, a su potencial de uso en investigación, gestión, toma de decisiones, y diseño de políticas ambientales. Al tratarse de una masa de datos en evolución, que crece continuamente, es importante señalar que los cambios observados son tan importantes como las estructuras detectadas, especialmente si marcan tendencias estables. Esto puede permitir pronosticar de qué se puede disponer en un momento dado, o qué inversión sería necesaria para corregir alguna tendencia o llegar a algún punto predeterminado, como un nivel concreto de conocimiento sobre alguno de los huecos de conocimiento que existan ahora.

En el estudio no se contemplan todos los factores de calidad, sino sólo algunos ejemplos que permiten apreciar la utilidad de las técnicas desarrolladas, y que pueden marcar líneas futuras de investigación. De este modo, puede considerarse este informe como un estudio piloto que puede facilitar la dirección de estudios más amplios, incluyendo un repertorio de técnicas. Una extensión evidente sería el estudio de la totalidad de los datos servidos desde proveedores españoles.

También, y quizás con mayor proyección, puede servir este estudio para establecer *mecanismos de monitorización* de la calidad de los datos servidos, que gbif.es puede desarrollar para contribuir a la evaluación de la calidad y la utilidad de sus datos. Estos mecanismos podrían, eventualmente, hacerse públicos a través de una herramienta de consulta que generase y sirviese los indicadores bajo demanda del usuario.

Metodología

En el análisis de la masa de datos hemos usado, en general, métodos similares a los que hemos empleado para analizar los datos totales de gbif.org, en el seno de una investigación general sobre pautas de datos que estamos realizando. Por lo general, se han aplicado las mismas técnicas pero restringiendo los resultados al proveedor gbif.es.

Estas técnicas buscan hacer emerger pautas espaciales, temporales o conceptuales en la base de datos, de forma que se puedan visualizar. Además de análisis clásicos de series temporales y pautas geográficas, hemos usado nuevas técnicas de representación que estamos desarrollando, como los cronhorogramas o los recogramas. En la mayoría de los casos, las pautas se han

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España
Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

buscado o detectado al representar en un espacio de dimensiones conceptuales (geográficas, temporales, taxonómicas) los datos agrupados de acuerdo a sus frecuencias.

Las técnicas empleadas han sido:

- Proyección de los datos sobre mapas según su localización geográfica, con representación temática o cuantificada de los indicadores de interés. Las frecuencias (o los grupos) de datos se representan en una escala de color sobre su ubicación geográfica marcada por sus coordenadas espaciales.
- Cronhorogramas de los datos, en los que se superponen pautas temporales lineales y cíclicas. Las frecuencias (o los grupos) de datos se representan en una escala de color sobre un mapa cuyas dimensiones, ortogonales o polares, representan los componentes temporales que se desea observar, y donde cada dato ocupa un lugar definido en esas coordenadas.
- Recogramas de los datos ("scary maps"), en los que los registros son representados secuencialmente según su momento de adquisición, en una escala de color según el atributo que interese.
- Representaciones proporcionales de frecuencias de grupos (treemaps) en los que el espacio se tesela con grupos jerarquizados, proporcionales al número de registros en cada categoría.
- Representaciones y estadísticas generales como tablas de contingencia, diagramas de barras, de áreas, etcétera.

Las técnicas se han aplicado mediante consultas directas SQL sobre las siguientes cuatro versiones del índice general de GBIF:

- Mayo 2008
- Octubre 2008
- Agosto 2009
- Septiembre 2009

que totalizan unos 3,5G registros de los que aproximadamente 650M registros son de datos primarios. Los datos se han filtrado a gbif.es, lo que da como resultado unos 3,6M de registros primarios en septiembre de 2009. Esto supone alrededor del 70% de los registros servidos por proveedores españoles, y el 2% del total mundial servido a través de gbif.org.

Algunos detalles de estas técnicas se comentarán al hilo de su uso en los siguientes resultados.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España
Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

Resultados

Reparto de los registros por su origen

Distribución temática

La mayoría de los registros se concentran en unos pocos datasets (Figura 1): Anthos (9090), Aves Reproductoras del INB (3366), Herbario de Plantas Vasculares del RJB (240) y Banco de Datos de la Biodiversidad de la Comunidad de Valencia (8004). Las plantas son el grupo dominante, con casi la mitad de los datasets.

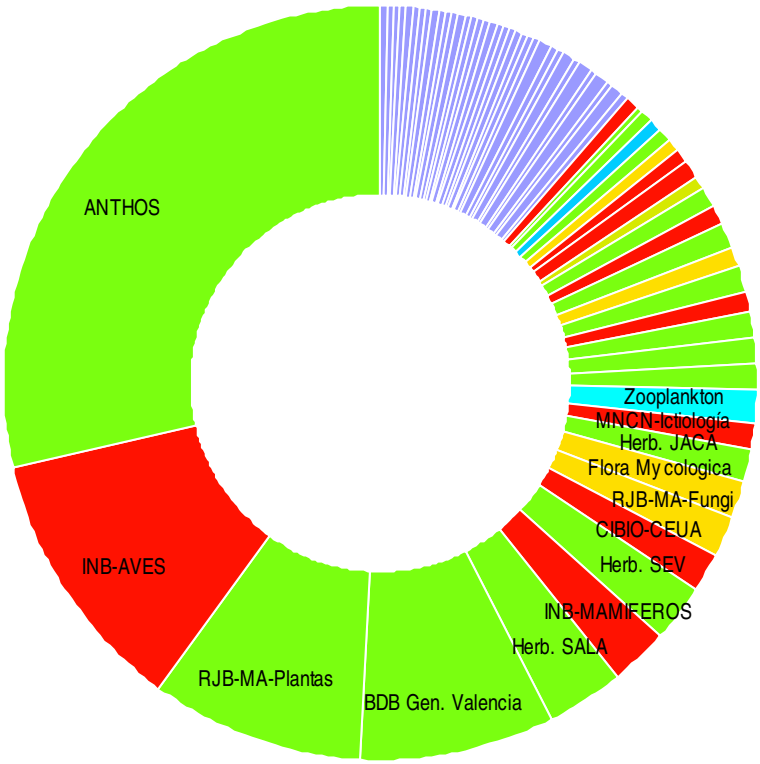


Figura 1: Distribución de los datasets en gbif.es (122 datasets). Sectores proporcionales al número de registros en cada dataset. Verde: plantas vasculares; rojo: vertebrados; amarillo: plantas no vasculares; azul: invertebrados. (En gris: sin asignar).

Naturaleza del registro

Por otra parte, la composición de los datasets y los registros no parece proporcionada en cuanto al tipo de dato que contienen (Fig. 2). Más de dos tercios de los datasets (78%) corresponden a bases de datos que contienen ejemplares, y un 19% contienen observaciones. Junto a estas, un 3% no declara el tipo de contenido o pueden ser, en realidad, agregadores o indexadores de otros proveedores. Pero la cantidad de registros de ejemplares es en realidad de menos de la mitad (43%); las observaciones son un 28% y hay una gran cantidad de datos (29%) sin asignar. La inmensa mayoría proceden de un único dataset: 9090, ANTHOS.

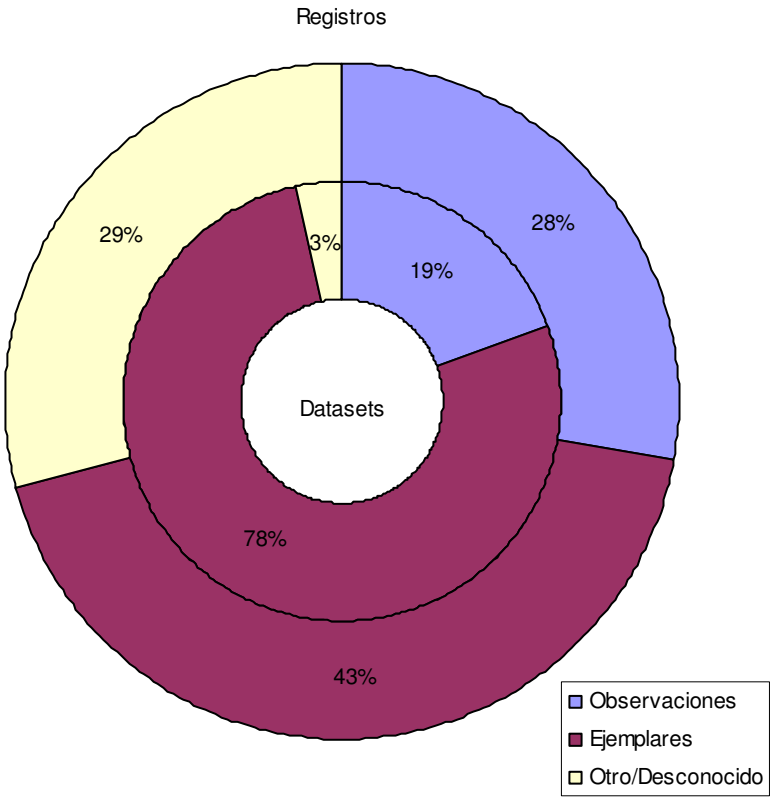


Figura 2: Tipo de datos en los datasets. Interior: número de datasets conteniendo cada tipo de datos: exterior: número de registros de cada tipo de dato.

Incorporación de registros

Este dataset (9090, ANTHOS) es el último que se ha incorporado, y ha alterado significativamente la distribución de registros. En la figura 3 puede observarse la secuencia de incorporación de registros en el índice, que crece de forma escalonada conforme se van incorporando aquéllos.

En la secuencia puede distinguirse con claridad un punto de inflexión en verano de 2008, con la incorporación de grandes datasets con datos de observaciones, además de la mencionada incorporación de 9090.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España

Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

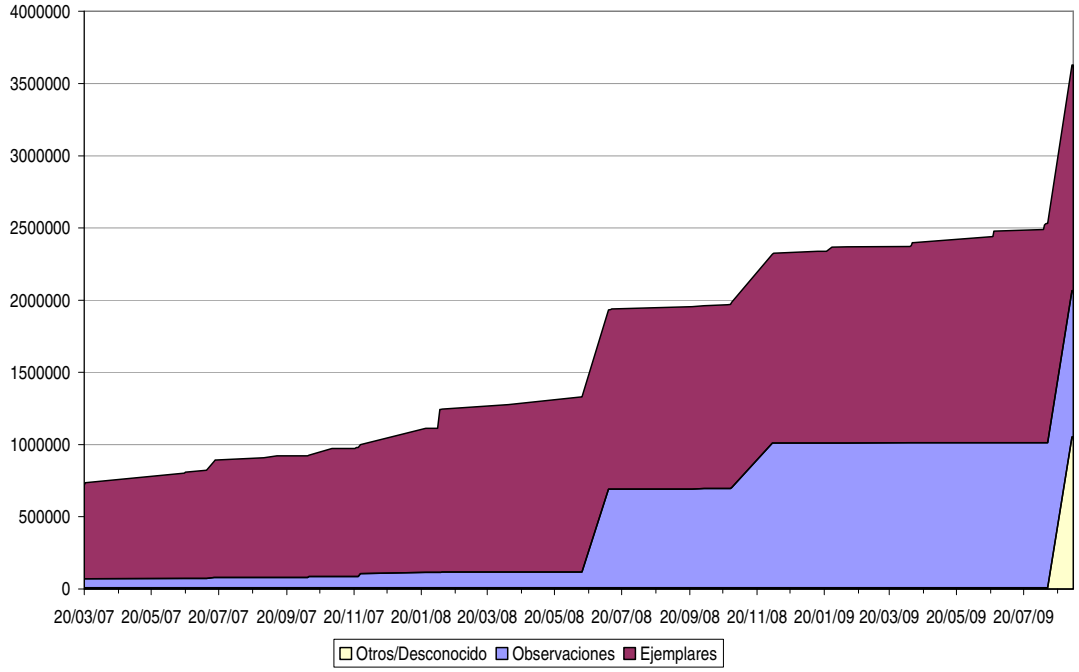


Figura 3: Incorporación de registros al índice en función del tipo de registro (Basis_of_Record).

Sin embargo, por lo general, las incorporaciones de registros al índice suelen ser eventos únicos (Fig. 4). Cada proveedor realiza un único volcado de registros, siendo relativamente infrecuente que se agreguen nuevos registros tras el volcado inicial; menos de la mitad de los proveedores realiza nuevas incorporaciones al índice.

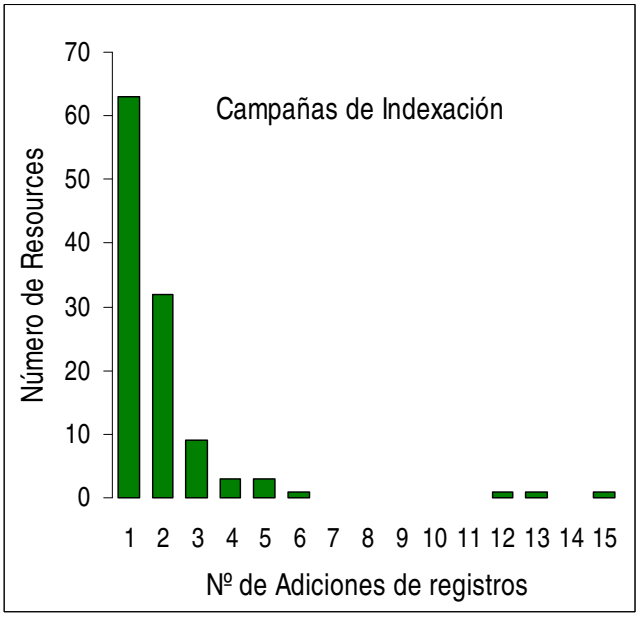


Figura 4: Campañas de indexación. Frecuencia de incorporación de nuevos datos a los datasets. La mayoría de los proveedores (60%) han realizado un único volcado.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España

Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

La incorporación de los registros al índice sigue la pauta taxonómica de los datasets, con algunos bloques homogéneos taxonómicamente (Fig. 5). Por lo tanto, los aumentos en el volumen de datos disponibles no necesariamente se traducen en un aumento proporcional en la cantidad de información disponible.

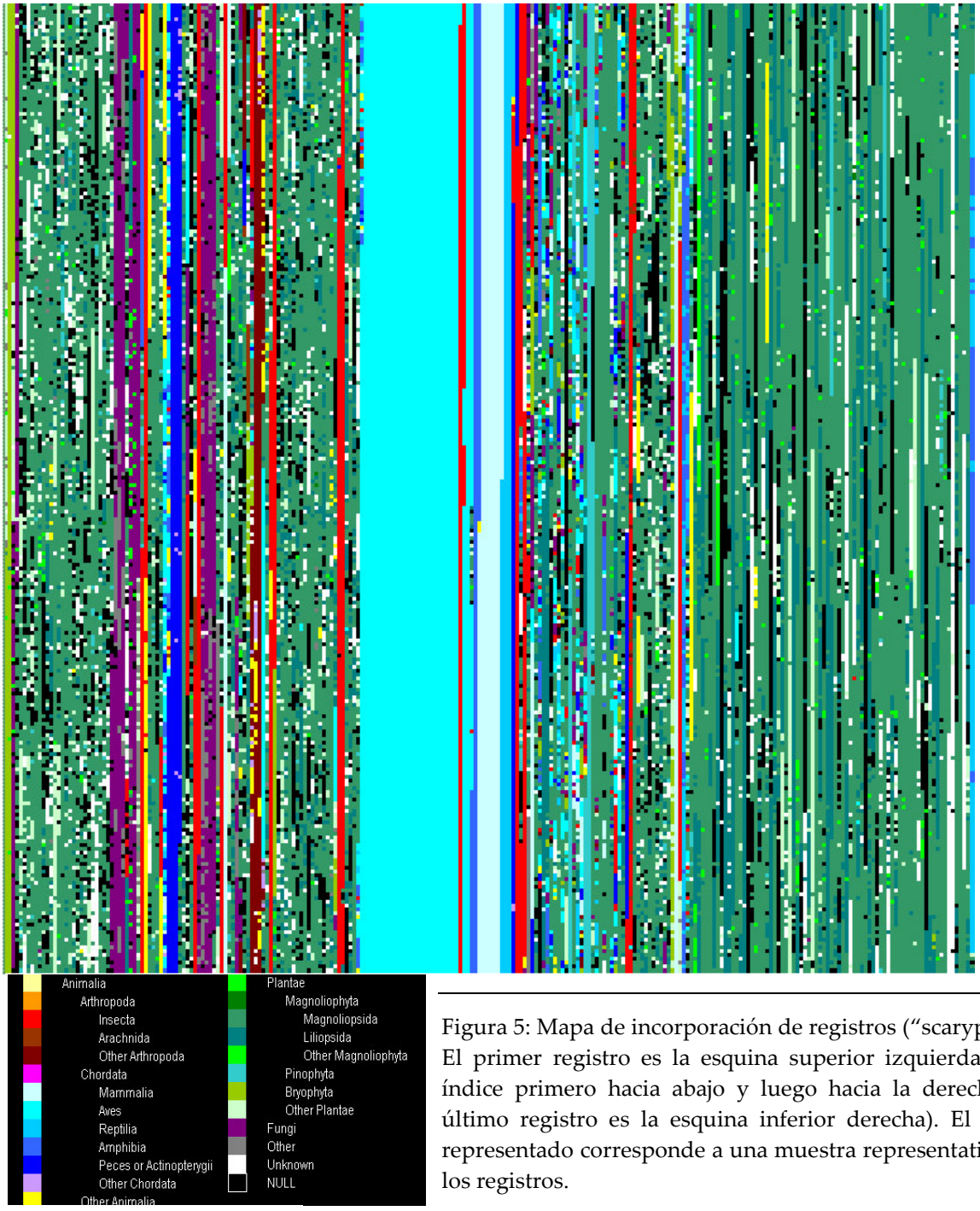


Figura 5: Mapa de incorporación de registros ("scaryplot"). El primer registro es la esquina superior izquierda, y el índice primero hacia abajo y luego hacia la derecha (el último registro es la esquina inferior derecha). El mapa representado corresponde a una muestra representativa de los registros.

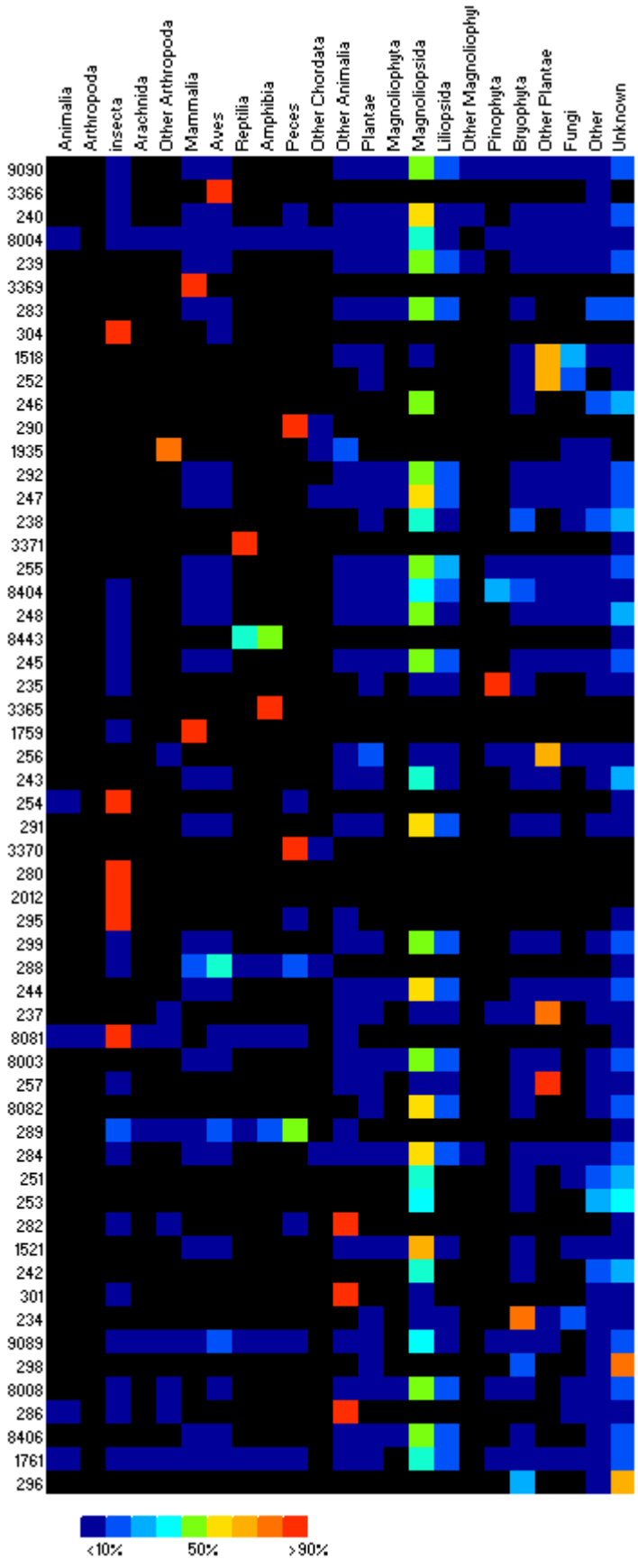
En un momento dato, la proporción disponible de información taxonómica dependerá de qué datasets se hayan incorporado. Por ejemplo, la información sobre Aves en los índices pasa de residual a extensa a partir de la incorporación de un proveedor concreto (bloque azul en el scaryplot de la figura 5).

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España

Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

Pautas taxonómicas

Distribución taxonómica de los datasets



Los datasets no son, en general, monotemáticos. Aunque en algunos datasets la información pertenece a un grupo taxonómico bien definido y en la mayoría hay una clara representación dominante en un grupo, en general se incluyen datos de otros grupos (Fig. 6).

Figura 6: Composición temática de la mitad más poblada de los datasets en gbif.es. Porcentaje de datos correspondientes a cada categoría taxonómica.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España

Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

Distribución taxonómica de los registros

La ubicación taxonómica se puede representar mediante treemaps jerárquicos (Fig. 7). En los registros, algo más de la mitad de los identificados corresponden al reino Plantae, pero hay un número alto de registros sin reino asignado. En los animales, hay un dominio claro de las Aves sobre los demás grupos, asociadas a las observaciones.

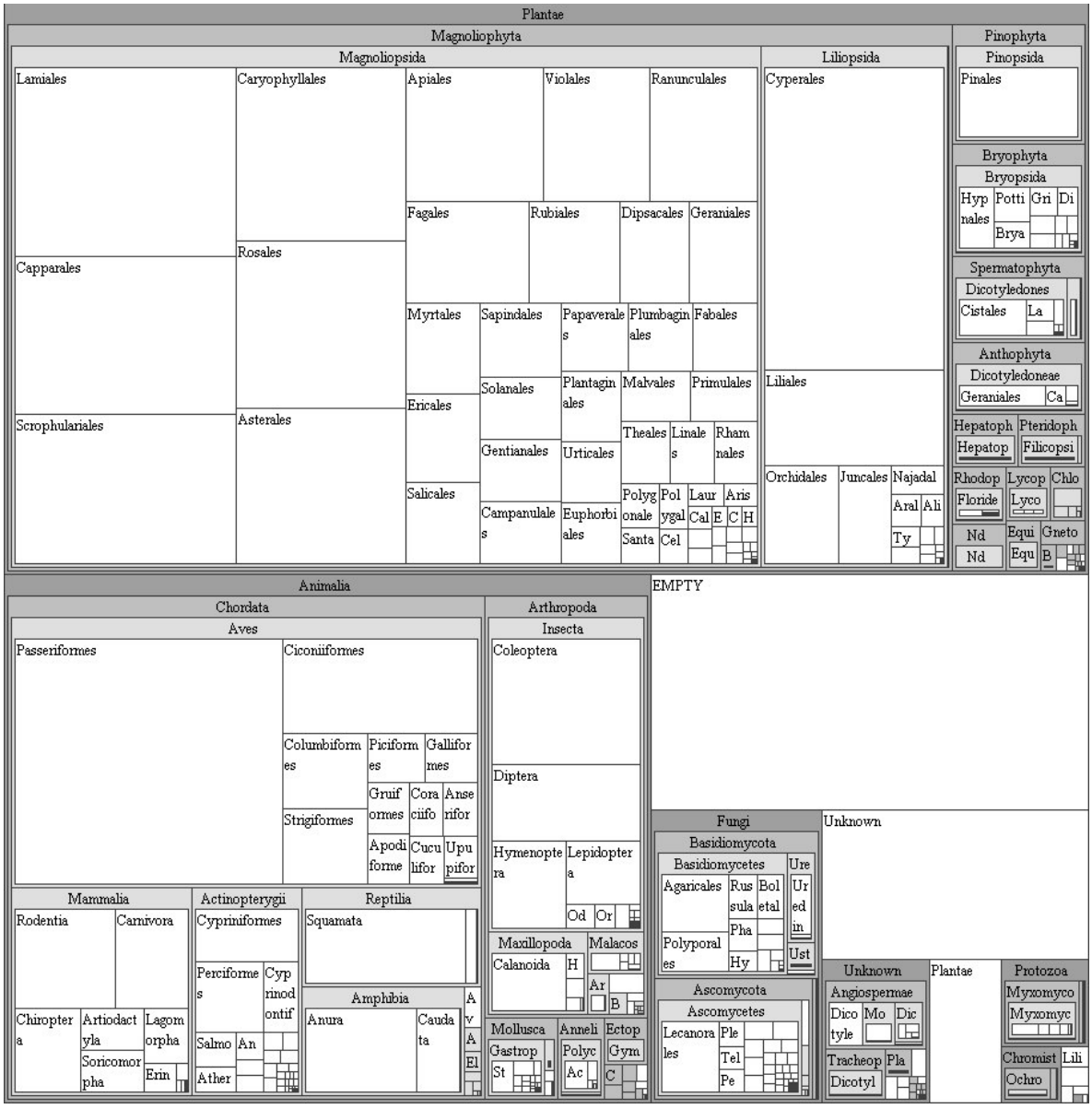


Figura 7: Treemap de la caracterización taxonómica de los registros. Las superficies de las teselas anidadas corresponden al volumen de registros en cada categoría.

Sin embargo, la clasificación anterior corresponde a la asignada durante la generación del índice, a partir del nombre de especie suministrado en el registro. Existen muchos registros para los que el proveedor facilita el nombre de especie, pero no su taxonomía o, si lo hace, la deja en niveles genéricos como Reino, lo que genera huecos en la taxonomía. La figura 8 representa el treemap de conceptos taxonómicos creados a partir de los datos de gbif.es.

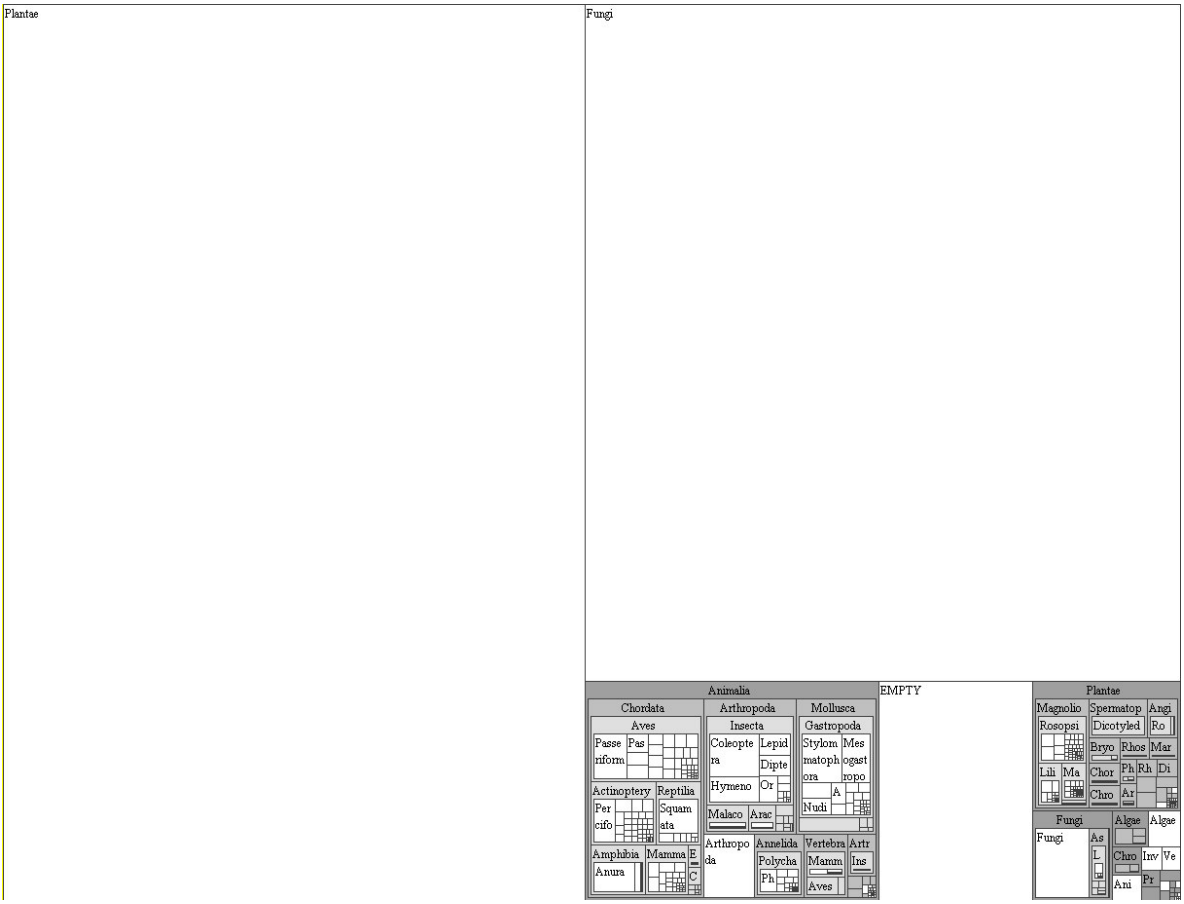


Figura 8: Treemap de la declaración taxonómica de los registros. Las superficies de las teselas anidadas corresponden al volumen de registros en cada categoría.

Además, la clasificación taxonómica suministrada por el proveedor es en general parca y fuertemente bimodal. Los registros o están completos en cuanto a taxonomía, o ésta falta por completo. La figura 9 muestra para cada dataset y para el total de datasets el porcentaje de registros en los que la taxonomía está completa (azul) o se ha declarado sólo hasta un determinado nivel (otros colores).

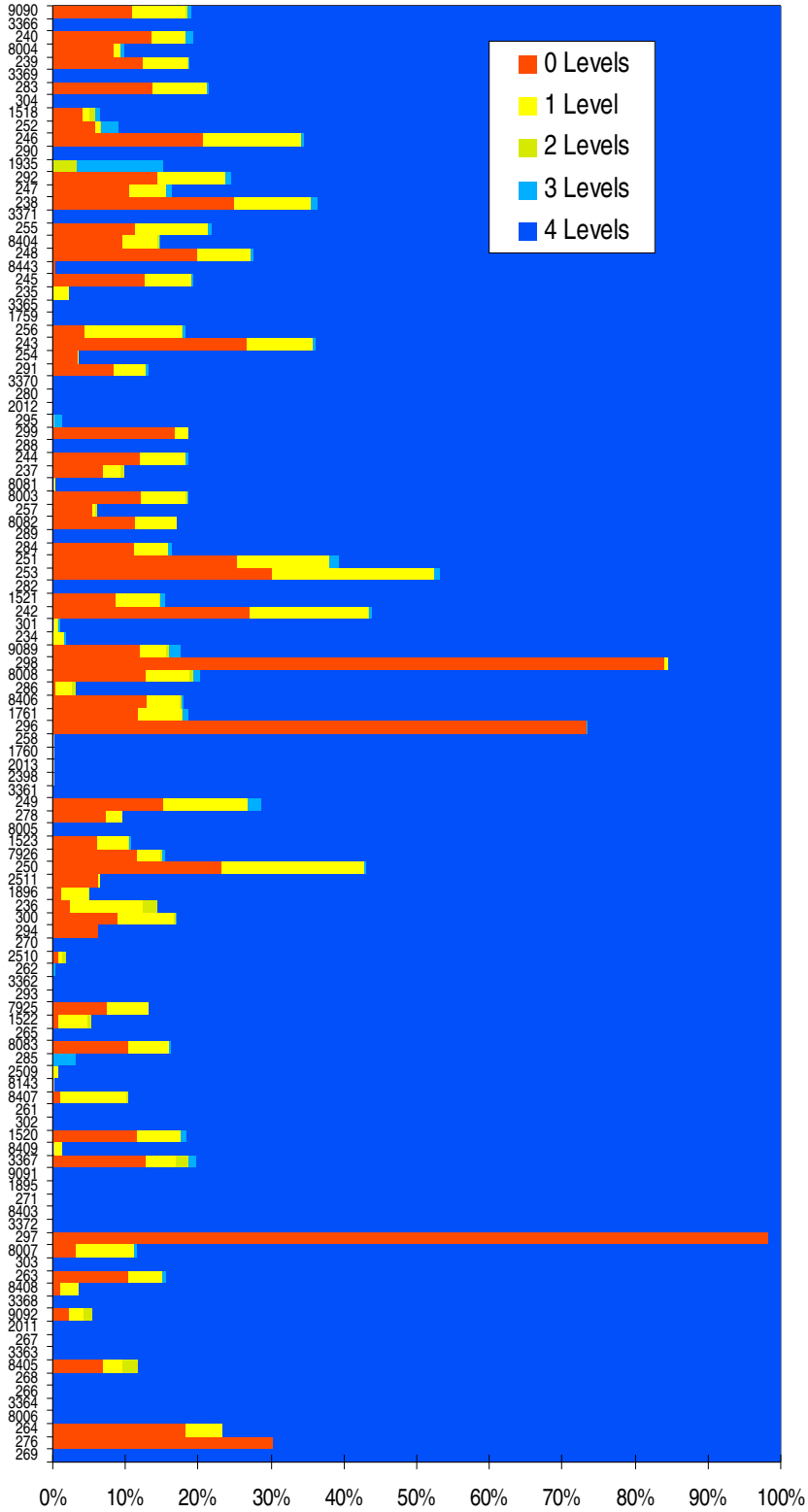


Figura 9: Porcentajes de registros de cada dataset en los que la taxonomía está completa (azul, cuatro niveles taxonómicos declarados: de reino a orden) o sólo ha alcanzado un determinado nivel. Los datasets (izquierda) están ordenados de más (arriba) a menos poblado.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España

Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

Pautas geográficas

La distribución geográfica de los registros que han sido georreferenciados se halla en la figura 10. Como era previsible, la gran mayoría de los registros corresponden a ocurrencias en España, pero se aprecian pautas de interés.

Existe una concentración de registros procedentes de lugares asociados a la geopolítica histórica española, como América del Sur, Caribe, Guinea Ecuatorial, Europa y el Norte de Africa. Pueden corresponder a registros procedentes de la época colonial, además de otros modernos de expediciones.

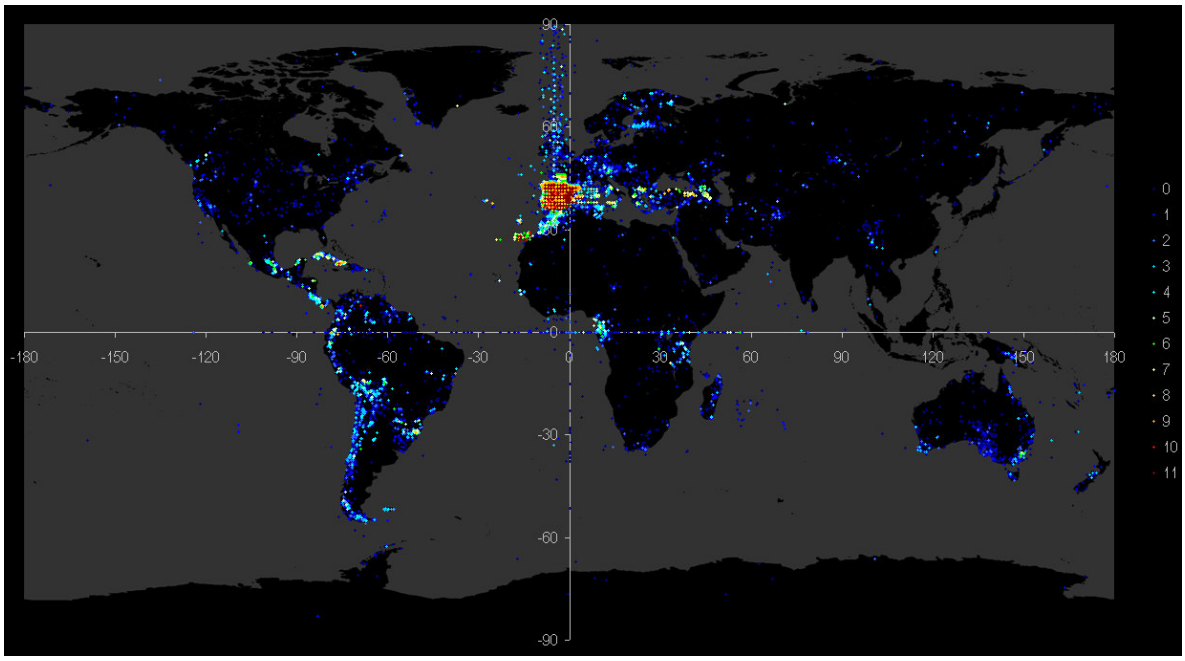


Figura 10: Localizaciones de los registros georreferenciados suministrados por gbif.es. La escala de color corresponde al logaritmo binario del número de registros que coinciden en las mismas coordenadas.

Algunas pautas parecen artefactos de georreferenciación, como la ristra que parte de la Península y sube hacia el norte. Los datos corresponden a un único dataset (283). Por tanto, esta representación permite detectar y corregir en este caso un probable error en las coordenadas.

A escala más local (Fig. 11) aparecen nuevas pautas. Las concentraciones de datos en la Comunidad Valenciana pueden asociarse a un único proveedor regionalizado. Las líneas horizontales en el Golfo de Vizcaya corresponden a datasets marinos, pero la malla regular de puntos en la Península y el Mediterráneo es evidentemente un artefacto de georreferenciación. Corresponde a registros cuyas coordenadas están redondeadas a grado, sin indicación de radio de incertidumbre. Por tanto, aparentemente localizan cada registro en la intersección lat-lon sin decimales, cuando la ubicación real podría ser cualquier punto con la misma referencia de grado.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España
Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

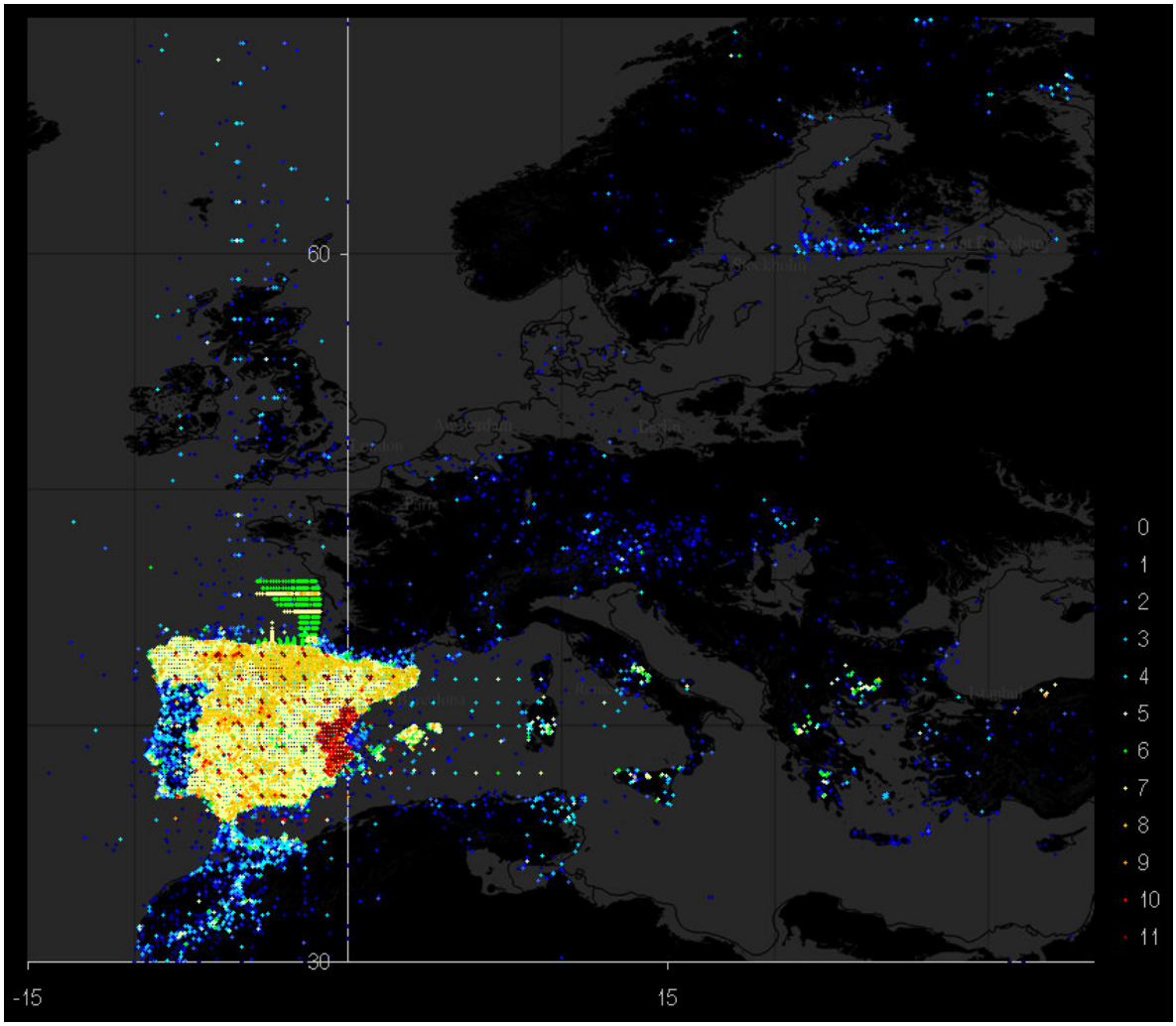


Figura 11: Localizaciones de los registros georreferenciados suministrados por gbif.es a escala europea. La escala de color corresponde al logaritmo binario del número de registros que coinciden en las mismas coordenadas. Las mallas regulares de puntos son artefactos de georreferenciación.

Otras pautas regulares pueden también indicar problemas en las coordenadas. Por ejemplo, el dataset más poblado (ANTHOS, 9090) suministra una georreferenciación que forma por lo general una malla regular, aunque está combinada son otras secuencias sorprendentes que deberían ser investigadas (Fig. 12a). Por su parte, el dataset 3369 (Fig. 12b) parece mostrar una pauta derivada de una conversión a partir de coordenadas UTM.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España

Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

Figura 12: Localizaciones de los registros georreferenciados suministrados por dos datasets dentro del proveedor gbif.es a escala peninsular. Escala de color como en la figura 11. Arriba, proveedor 9090. Abajo, proveedor 3369.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España

Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

La distribución de la georreferenciación en los datasets es fuertemente bimodal (Fig. 13): dentro de un dataset, lo común es que los registros o no estén georreferenciados o lo estén todos; es el caso de más de la mitad de los datasets.

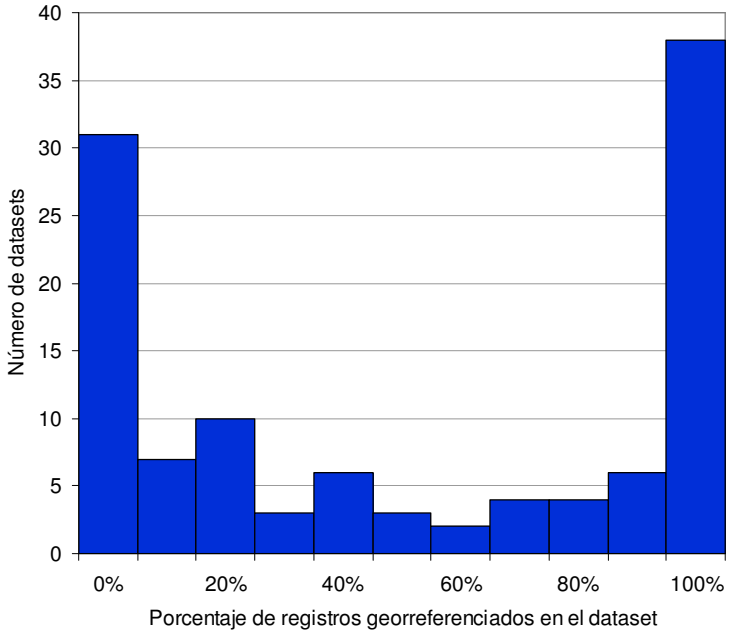


Figura 13: Histograma de la frecuencia de georreferenciación. Porcentajes de registros con algún nivel de georreferenciación en cada dataset.

Pautas temporales

Situación en el tiempo de los datos

Los datasets servidor por gbif.es contienen datos que abarcan tres siglos, aunque naturalmente concentrados en tiempos recientes (Fig. 14). La mayoría de los datos corresponden a ocurrencias ubicadas en el último tercio del siglo XX, disminuyendo en ambas direcciones. Esto puede ser indicativo de varios posibles fenómenos:

- Incorporación de datasets históricos;
- Dificultad de incorporación de datasets correspondientes a investigaciones recientes (¿"efecto embargo"?);
- Incorporación lograda de la mayoría de los datasets ya en formato digital (efecto "low-hanging fruit");
- Deriva hacia otros proveedores por generación de capacidad.

En la secuencia puede también apreciarse con claridad el efecto de eventos puntuales significativos desde el punto de vista de la generación de datos, como la guerra civil.

El conocimiento de la distribución temporal de los datos puede ayudar a establecer hasta dónde es posible establecer *baselines* para estudios evolutivos que requieran datos de biodiversidad; las *baselines* pueden verse fuertemente apalancadas por la concentración de datos en ciertos momentos.

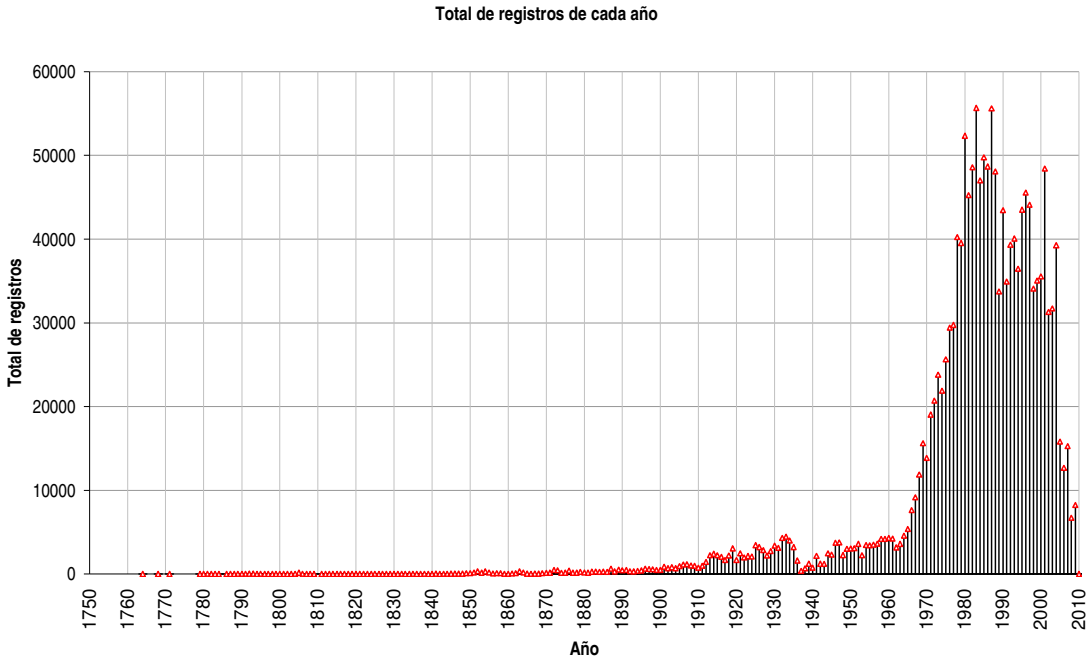


Figura 14: Número de registros fechados cada año en la serie suministrada por gbif.es. (Los registros sin dato de fecha no aparecen aquí).

Como en el caso de la georreferenciación, la distribución de fechas en los datasets es bimodal, aunque en este caso hay una clara mayoría de datasets que suministran todos sus registros con fecha. Sin embargo, cerca del 15% de los datasets están suministrando registros que no tienen fecha (Fig. 15).

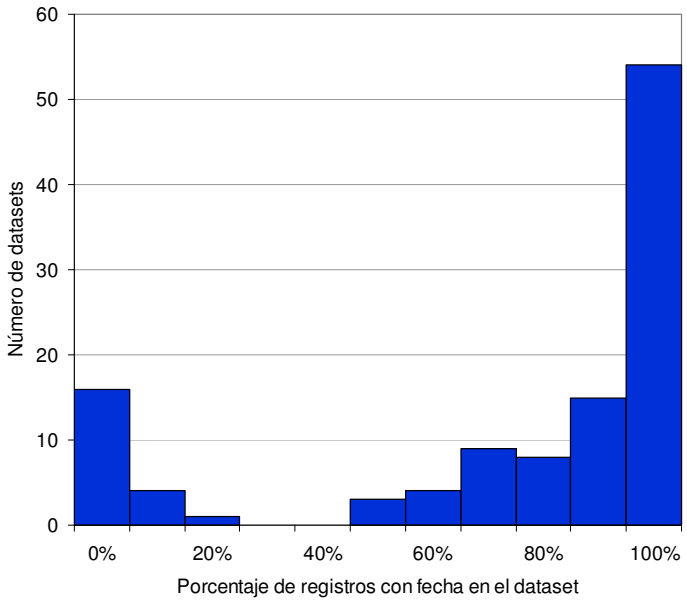


Figura 15: Histograma de la frecuencia de datación. Porcentajes de registros con algún nivel de datación en cada dataset.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España

Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

Pautas estacionales

La distribución de las fechas de los registros parece también obedecer a una pauta cíclica, relacionada con la estacionalidad y superpuesta a la pauta anual En la figura 16 se representa el cronhorograma de los registros a partir del siglo XX, donde puede apreciarse una mayor concentración de datos en los meses de verano (lado izquierdo) así como los huecos importantes de datos, como la guerra civil (anillo oscuro).

Además, pueden verse errores evidentes de datación. Existen algunos datos (puntos fuera del círculo principal) correspondientes a fechas futuras. En general, estas fechas exóticas pueden atribuerse a datasets muy concretos que pueden contener errores sistemáticos de fecha. Por ejemplo, los datasets 240, 1518 y 8008 presentan fechas que llegan al año 7000.

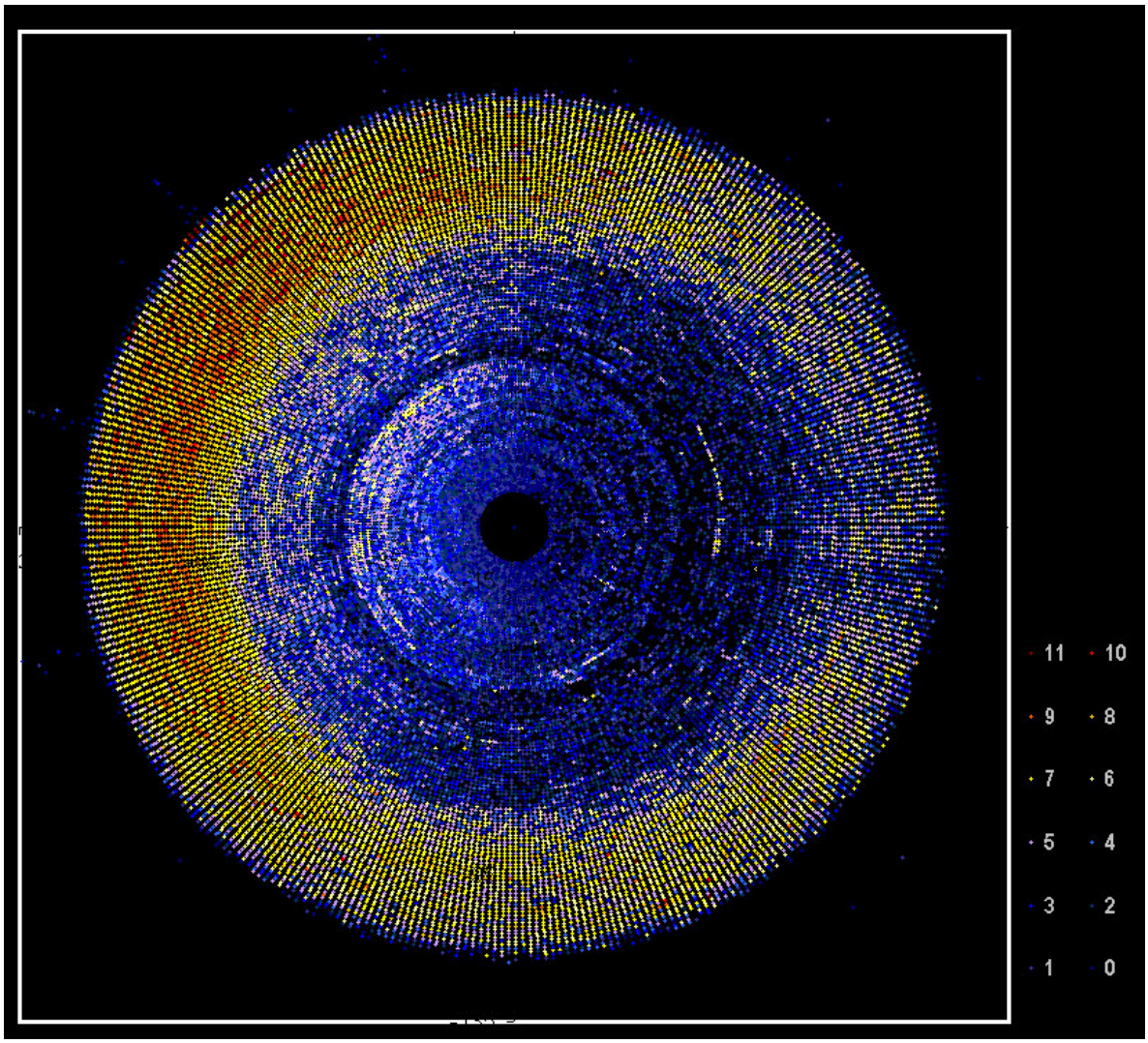


Figura 16: Cronhorograma de los registros en los datasets servidos por gbif.es. Cada punto representa el número de registros datados en un día; la escala de color es el logaritmo binario de ese número. Las coordenadas son polares. El radio representa el año, comenzando en 1900 (centro), y el día del año está representado a lo largo del círculo, comenzando en la posición de las tres en punto (día primero, 1 de enero) y progresando en el sentido contrario a las agujas del reloj.

La pauta global estacional puede observarse en la figura 17, en la que se indica el promedio de registros fechados en cada día del calendario.

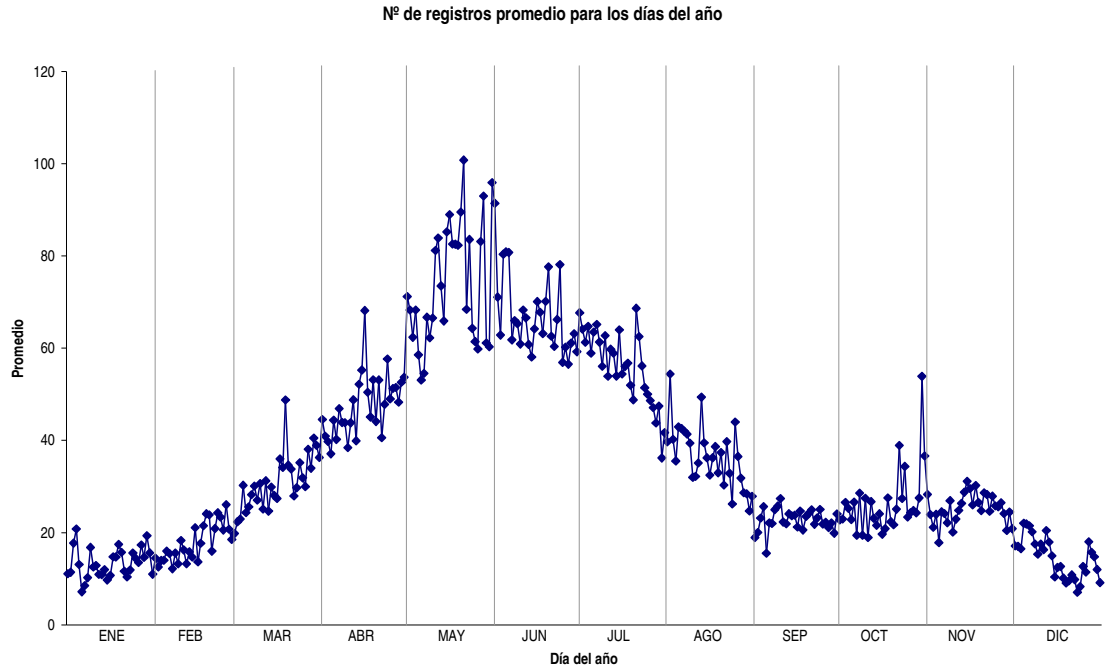


Figura 17: Número medio de registros datados en cada día del año a lo largo de todas las series.

Asimismo, existen otras pautas cíclicas de otros períodos, como pautas mensuales y semanales (Figura 18). Por ejemplo, parece darse una concentración importante de datos en los fines de semana de verano. Esto podría estar relacionado con el aporte de observaciones a agregadores por parte de los ahora llamados "citizen scientists", personas privadas que realizan observaciones en su tiempo libre y las comunican a algún servicio de recogida de datos, como en el caso de las observaciones de aves. Sin embargo, la pauta que parece apuntar al primer día del mes podría ser artefactual: una proporción anómalamente alta de registros con día=1 (por encima de la fracción de días 1 en el año) podría indicar simplemente una imprecisión en la datación.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España
Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

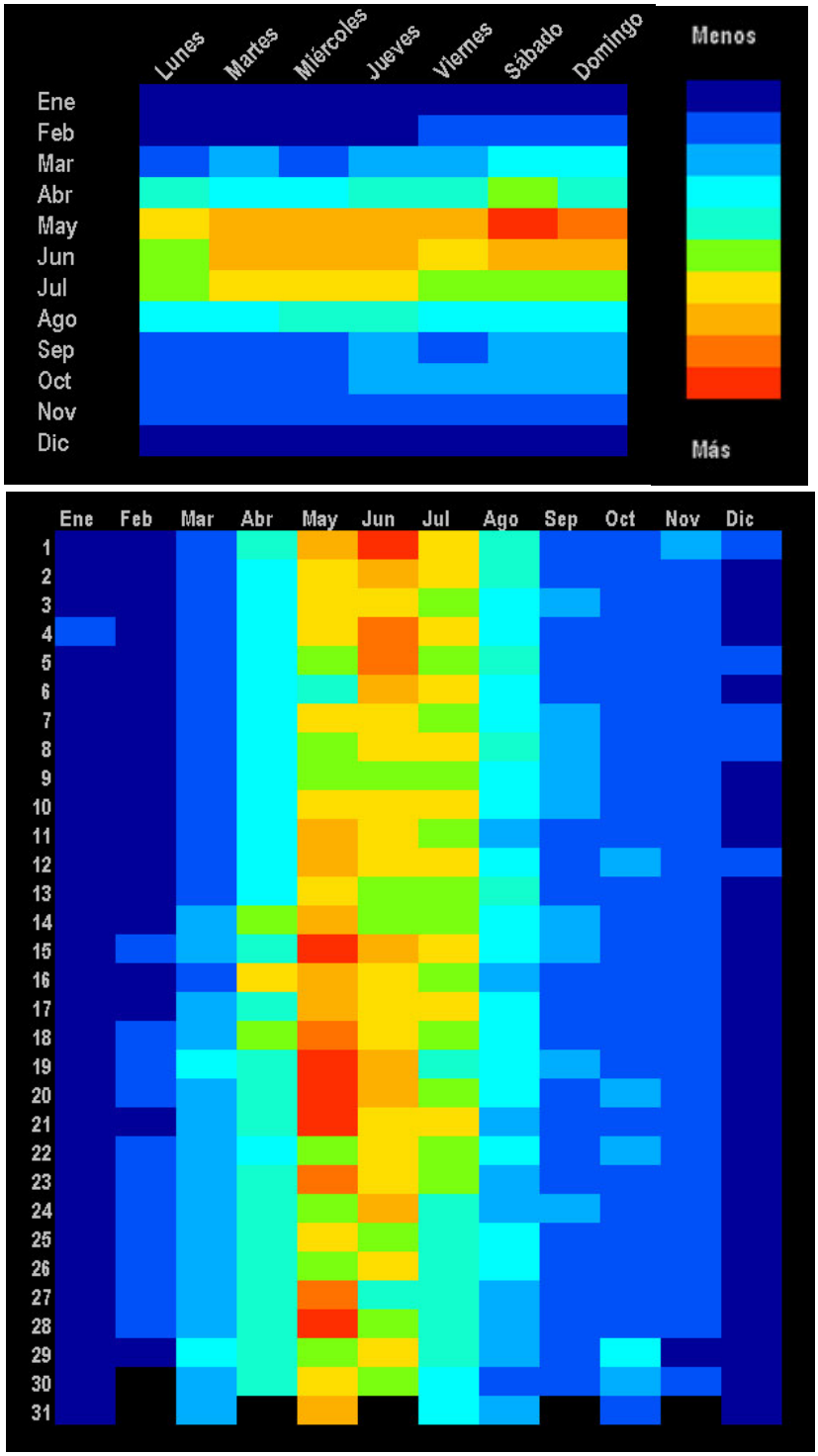


Figura 18: Frecuencia relativa de registros según el día de la semana y el mes (arriba) y del día del mes y el mes (abajo).

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España

Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

Si bien en el conjunto de los registros no parecen observarse problemas artefactuales importantes excepto por el primer día de algún mes, como se ha mencionado arriba, si los datos se examinan para cada dataset sí pueden detectarse algunos datasets (poco poblados) en los que parece apuntarse algún problema de fechas. La mayoría de los problemas están relacionados con anotaciones "virtuales", como poner día 1 (o mes 1) para indicar cualquier día del período correspondiente. En la figura 19 se representan todos los datasets ordenados de acuerdo al porcentaje de registros que contienen en los que la fecha está anotada como el día 1 de cualquier mes o el día 1 del año. Si los registros procedieran de una muestra aleatoria de los días del año, aun cuando sufrieran una concentración estacional, las proporciones de días 1 del mes o de día 1 del año deberían aproximarse a alrededor de 1/30 y 1/365, respectivamente, lo que no sucede en un número importante de datasets.

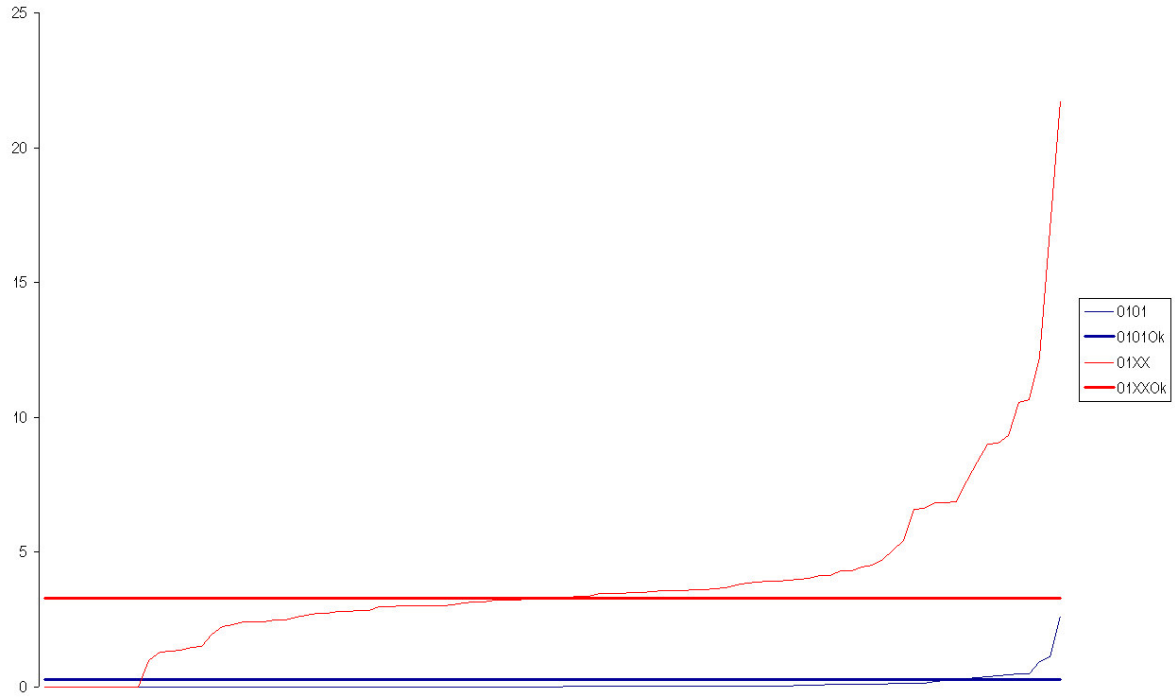


Figura 19: Proporción de registros (ordenadas) dentro de cada dataset (abscisas) correspondientes al primer día del mes (rojo) o del año (azul). Las líneas horizontales representan las proporciones ideales de tales días en un dataset cualquiera. Los datasets de la derecha (no son los mismos para ambas curvas) tienen un número anómalamente alto de registros datados en días 1, y los de la izquierda un número anómalamente bajo.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España
Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

Evolución de la calidad

La comparación de cuatro versiones del índice separadas por 18 meses permite observar cómo ha ido evolucionando la calidad de los datos servidos en cuanto a algunas de las pautas anómalas detectadas.

En la figura 20 se indican los porcentajes de registros existentes en cuatro ediciones del índice que tienen datación o georreferenciación. Puede verse cómo la proporción de registros que estaban georreferenciados ha ido aumentando a lo largo de las ediciones, pero en cambio la proporción de registros datados ha ido disminuyendo.

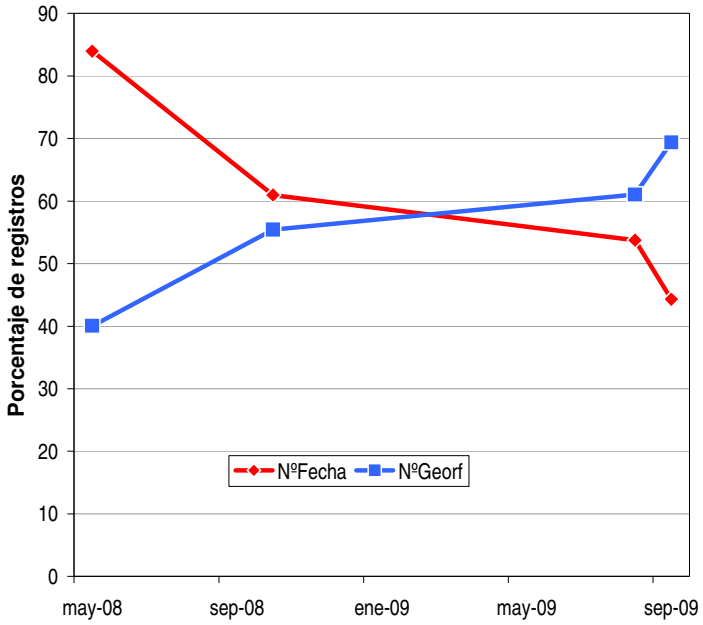


Figura 20: Proporción de registros georreferenciados (azul) o datados (rojo) existentes en cuatro versiones del índice GBIF.

Atendiendo a la tasa de errores explicados más arriba, la figura 21 muestra la evolución de algunos casos significativos: coordenadas imprecisas en la georreferenciación, y fechas sospechosas. La proporción de problemas de georreferenciación parece estar reduciéndose, al hacerlo (entre otras cosas) la fracción de registros en los que las coordenadas están dadas sin decimales, que resultan ambiguas en ausencia de un indicador de radio de incertidumbre. También disminuyen los registros con coordenadas 0,0. Sin embargo, la proporción de registros con fecha 1 de cualquier mes se mantiene constante, así como (aunque en porcentajes muy bajos, nótese la escala logarítmica) la de fechas con otros problemas. Las fechas futuras se han ido perdiendo, pero las fechas muy antiguas o de 1 de enero parecen estar volviendo a repuntar tras una caída más o menos continua. Esto podría estar relacionado con las últimas incorporaciones de datasets.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España
Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es

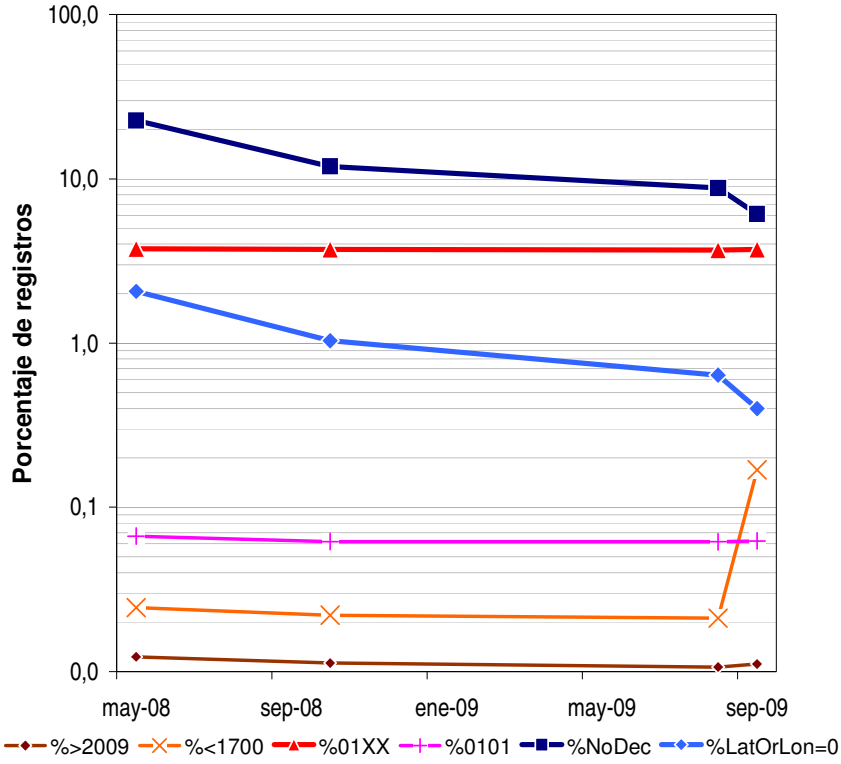


Figura 21: Proporción de registros georreferenciados (azul) o datados (rojo) existentes en cuatro versiones del índice GBIF.

La conclusión final acerca de la evolución de la calidad de los datos atendiendo a estos dos parámetros sería que está mejorando la georreferenciación, pero en cambio está habiendo problemas crecientes de datación en los datasets.

Universidad de Navarra, Pamplona, 17 de diciembre de 2009.

Arturo H. Ariño *Javier Otegui*

La matriz digital de este documento se sella permanentemente con la clave pública correspondiente a PGPID:0xFE08ED42 a las 18.07 del 17/12/2009.

Dr. Arturo H. Ariño - Departamento de Zoología y Ecología, Facultad de Ciencias, Universidad de Navarra, Irunlarrea s/n, E-31080 Pamplona, España

Teléfono: 948-425600 Fax 948-425649 Télex 37917 UNAV E Internet artarip@unav.es