# ON THE DATES OF THE GBIF MOBILISED PRIMARY BIODIVERSITY DATA RECORDS

JAVIER OTEGUI (1), ARTURO H. ARIÑO (1)*, VISHWAS CHAVAN (2) AND SAMY GAIJI (2)
*(1) Department of Zoology and Ecology, University of Navarra, Pamplona, Spain.*
*(2) Global Biodiversity Information Facility Secretariat, Universitetsparken 15, 2100, Copenhagen, Denmark.*
**corresponding author*

*Abstract* — There are more than 390 million primary biodiversity data records published by hundreds of data publishers through the GBIF network. Thus, the GBIF network is the single most comprehensive index for this kind of data. Ensuring or, at least assessing data quality is of capital importance for the reliability and usability of this data. While conducting a time data gap analysis on this mass of data, we have detected some issues with the way date information is processed and shared. Dates can be obscured or altered under certain circumstances, when a specific combination of publisher's error or date handling features, and faulty or inadequate date parsing and processing routines gets chained together. The extent of the date unreliability (either at the source or through GBIF portal) is relatively low, and problems are concentrated in a few data publishers. The types of errors and misprocessing in dates through the sources and the published records are analysed, impact on the overall data quality of the published index was assessed, and corrective measures are suggested.

*Keywords* — Primary biodiversity data, GBIF, dates, data quality, fitness for use.

The Global Biodiversity Information Facility (GBIF) is the inter-governmental organization aimed at ensuring free and open access to the world's primary biodiversity data (GBIF 2008). Two aspects are key to this effort. First, the ability to provide a common, consistent informatics infrastructure and schema that can collect information from disparate sources: as these may have their records in different formats and under different schemata, information exchange standards such as the Access to Biological Collections Data (ABCD) (TDWG, 2007) and DarwinCore (Darwin Core Task Group, 2009), complemented by retrieval tools such as the TDWG Access Protocol for Information Retrieval (Tapir) (TDWG, 2010) funnel the information into a common data model. Second, the development of simple yet powerful tools such as the GBIF data portal (http://data.gbif.org), that may serve as a primary discovery tool and allow efficient indexed data extraction and are amenable for easy query through the portal.

Joining data records published by the data publishers into this single data model entails a large amount of data mapping, checking, marking and manipulation. Concerns about the coverage and the quality of the data being made available through the GBIF network, both through the available web services or its data portal after indexing, have been expressed by several key stakeholders (Chapman 2004, Yesson et al. 2007, Boakes et al. 2010, Hill et al. 2010). This has prompted a number of initiatives to investigate into overall fitness for use status of the data being made available through the GBIF network. In recent past, the GBIF Secretariat and the University of Navarra tried to assess the taxonomic, geospatial, temporal coverage and quality of the GBIF mobilised data using different approaches (Gaiji et. al., 2013, this volume).

The overall time-related data gaps were described in aforementioned paper (Gaiji et. al., 2013, this volume) along with spatial, taxonomical, and other gaps. However, in the course of the analyses a number of issues specifically related to the date information emerged from the comparison of the results obtained by both groups. In this paper we describe in further detail these issues, try to gauge to what extent they can impact the quality of mobilised data, and suggest corrective measures.

## DATES IN GBIF DATA CACHE

From the data made available to the users by each data publisher, GBIF collates a data cache (hereinafter "index") that ensures the traceability of the data published by the data publishers. This index is essential to ensure seamless, easy and efficient discovery and access to data at the multiple sources provided by the separate data publishers. Such an index is based on the existence of two large tables. The first table (hereinafter

"RAW table") holds the verbatim information released by data publishers, which is cached "as it is" whenever possible. The second table (hereinafter "OCC table") caches the occurrence records once mapped to a common data structure and processed through a complex set of filters. The filters interpret the contents in RAW, look for inconsistencies and, ideally, deal with all suspicious information (such as out of range coordinates or invalid date values) by nullifying or flagging the corresponding record. OCC table has fields optimized for consistent, fast search and retrieval of data. The flow of data goes thus from the RAW table to the OCC table passing through the filters (Fig. 1), and it is the OCC table (acting as an index for RAW and, therefore, the publishers' data), that is exposed to queries through the GBIF portal. The processing is done cyclically at the GBIF headquarters in Copenhagen upon harvesting the data from the publishers. As of the end of 2010, the full cycle (including harvesting the data being made available by the publishers) took about six weeks on average. Once the full process is finished, the new index replaces the old one at the GBIF portal, is also deployed at the GBIF mirrors distributed across the world, and a new harvesting/processing cycle starts.

The temporal data string for occurrence record is thus available in both tables. In the RAW data table that has been populated with records collected from the data publishers at the time of caching, three fields represent the three levels of every date of the occurrence records: *day*, *month* and *year*. As these fields are collected verbatim from the data publishers, they may or may not represent a valid date in any given standard format. For instance, the field for month (defined in DarwinCore as part of a date which should use the ISO 8601 encoding scheme, as specified in the reference guide available at http://code.google.com/p/darwincore/wiki/Event), can actually hold any number, not restricted to the range 1-12 (or 0-11 for some data publishers; Tim Robertson, pers. comm.), or even character string values ('Jul.' is an example). Also, fields can contain zero or null values (this difference matters, as we will discuss later), to represent missing data. Some pre-processing takes place to atomize dates from certain publishers using ABCD schema (Tim Robertson, pers. comm.)

In the OCC table that is produced from the raw data, there are three fields for *year*, *month*, and *full date* (but not *day*). Month and year are interpreted from the equivalent fields in the RAW table, whose values they ideally match if valid, while invalid values in RAW become nullified in OCC. The full date field, in turn, is constructed by interpreting the year, month and day fields in RAW as best as possible. If the reconstructed date is consistent (*i.e.* represents a valid date), the field is filled with this value; otherwise, the field gets nullified (fig. 1).

### ADVANTAGES AND DISADVANTAGES OF THE TWO-WAY INDEXING OF DATES

This two-way caching of dates in the OCC table may facilitate fast retrieval and filtering of dates at different levels of granularity. As most time-filtered searches would conceivably be segmented by year (time-series targeted research) or month (seasonal targeted research), maintaining a separate field for these data helps avoiding their extraction from a full date value for every record that is searched through, while simultaneously maintaining the full date string as interpreted from the source. In terms of caching efficiency, this seems a good compromise, ideally leading to a better performance than having either all information only in one field, or split in three.

Nevertheless, it has a potential disadvantage. As the full date field in OCC is interpreted from the three separate fields in RAW, inconsistencies should in principle be limited to nullified date fields while year and/or month may exist separately. Year and month within the full date field, and the corresponding year and month in their respective fields, should be consistent within OCC table and should also match existing RAW values. However, RAW values need to be interpreted whenever these do not result in a null date. Should any inconsistency result from that interpretation other than nullification of full dates (for example, differences in the year field in RAW and/or OCC, and the corresponding year portion of the full date field in OCC), doubts would be cast on what would be the actual occurrence date, jeopardizing the quality of the data.

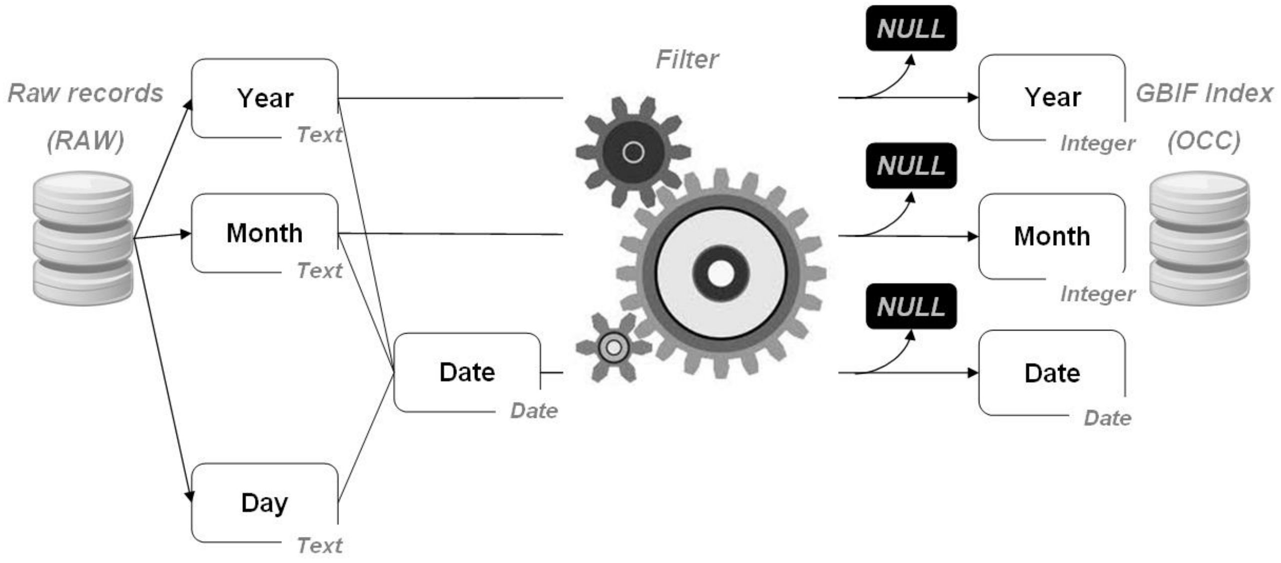


Figure 1: A segment of date processing at GBIF. RAW table contains verbatim information as provided by data publishers after pre-processing during harvesting (Robertson, pers. comm.); OCC contains ready-to-index parsed information. If any of the fields contains uncorrectable, non-valid information, the filter would, ideally, exclude it from publishing.

## METHODS

While attempting to evaluate the extent of the date gaps in the index, inconsistencies were found. The inconsistencies were investigated, aiming first to detect what types of discrepancies between RAW and OCC caused by inconsistencies may have resulted; second, to elucidate their possible causes; third, to determine whether these may significantly impact data gap analysis; and finally, to assess the scope of these issues and to what extent they might be affecting the fitness-for-use of the temporal data made available by publishers through GBIF's index.

Differences on individual records' reported dates existed between the original RAW and the OCC tables, either intended or unintended, likely caused by the filters' parsing algorithms treating specific date field contents. We classified these differences into "good", accomplishing the aim of improving data quality, and "bad", degrading the quality of records in the searchable index published through the portal. Note that although the user of the portal can navigate to the original, published records, dates searched are done by querying the OCC table.

"Good" combinations include:

- RAW data null and OCC data null. If data is not provided in the original set, no information should appear once processed.
- RAW data valid and OCC data valid and matching. The information a priori valid within the RAW dataset should be maintained in OCC.
- RAW data invalid and OCC data null. If the information contained in RAW is not valid, the filter should detect it and block its way to the index, or at least flag it as invalid.

"Bad" combinations include:

- RAW data null and OCC data valid. If the original RAW record fields have no information, OCC should not make up this information in its corresponding fields.
- RAW data valid and OCC data null. Blocking valid records from OCC reduces the availability of temporal information.
- RAW data valid and OCC data invalid, not null. This transformation has the same effect as nullifying valid RAW records, as false dates cannot be retrieved.
- RAW data invalid and OCC data invalid, either matching or not. The only valid process to do with invalid data is to nullify it when passing from RAW to OCC tables. Any other

action (including the lack of any action) would mean a reduction of quality of the dataset.

In addition, there are two combinations which may be either "good" or "bad":

- RAW data valid and OCC data valid but not matching. If valid, real information is altered, false information may result and the real data in RAW cannot be found. However, valid, unreal data can exist in RAW that can be corrected into valid, real data in OCC when day and month fields in RAW are swapped. If both month and day fields in RAW are below 12, there is no way to tell whether they are correct or not, unless by comparison with other records from the same data publisher. Even so, datasets may have certain records reversed while others are not.
- RAW data invalid and OCC data valid. When the sources of error in RAW date tagging are known and remediation exists (i.e. solving a date/month swap, or correcting outliers in year fields), there is an improvement in the availability of data. However, this transformation may also result in false dates if the cause for invalid RAW data is incorrectly identified or treated.

### *Data mapping and methodology for the analysis*

We examined the November 2010 full version of the GBIF index. We extracted both RAW and OCC tables and compared them record-by-record using MySQL scripts. Records were matched by using their unique identification number, assigned to the record in RAW at the time of harvesting and maintained in the OCC table by GBIF.

We classified each time field in each record in both RAW and OCC tables as belonging to one of three categories: null, non-valid, or valid. We designated as non-valid any day outside the range 1-31, any month outside the range 1-12, and any year outside the range 1750-2010. Dates in OCC table were designated as non-valid if they had any component designated as non-valid, and as null if either month or year field was null. We recorded discrepancies in both the category each component in the time fields belonged to, and in the numerical differences in valid dates between RAW and OCC, and between the two atomised fields in OCC (year and month) and the corresponding year and month in the full date field.

After compiling a table of all types of discrepancies between categories, we searched for patterns in the statistical distribution of discrepancies using categorical plots and time series plots, trying to figure out any regularity that might hint to a processing problem or a feature on the source data themselves.

## RESULTS AND DISCUSSION

There were 269,297,636 records in RAW table and 267,380,680 records in OCC table, with more than 99% of records (267,233,796) common to both tables.

Out of all the common records, 172,322,713 (64%) had information in all three time fields in RAW, but only 122,406,532 records (46%) in OCC table had retained non-null information in their date field. By contrast, the field 'year' was more complete: 73% of common records (196,085,880 in RAW and 196,015,248 in OCC, less than 0.04% difference) had the field filled (Gaiji et al., 2013, this volume).

A record in OCC can contain a non-valid, non-null value only if the year is out of range, as the field is of date type and becomes automatically null if an invalid or incomplete date is passed. However, in the RAW table any combination in the three time fields is possible, thus potentially yielding invalid dates even if no date field is null. Figure 2 shows the distribution of 12,084,317 records (4.5%) in RAW having at least one non-valid time field, and the distribution of 94,911,083 records (35,5%) in RAW having at least one null time field, according to the offending field(s) and their combinations. Three-quarters of the non-valid dates were caused by both day and month being invalid, while year problems were present in only about 12% of records having issues. Also in three-quarters of all nullified records, the reason for nullifying was lying in the three date levels simultaneously (year, month and day, 74.5% of all cases), while most of the remaining nullified records (22%) were so because of day and month being invalid, while retaining year information.

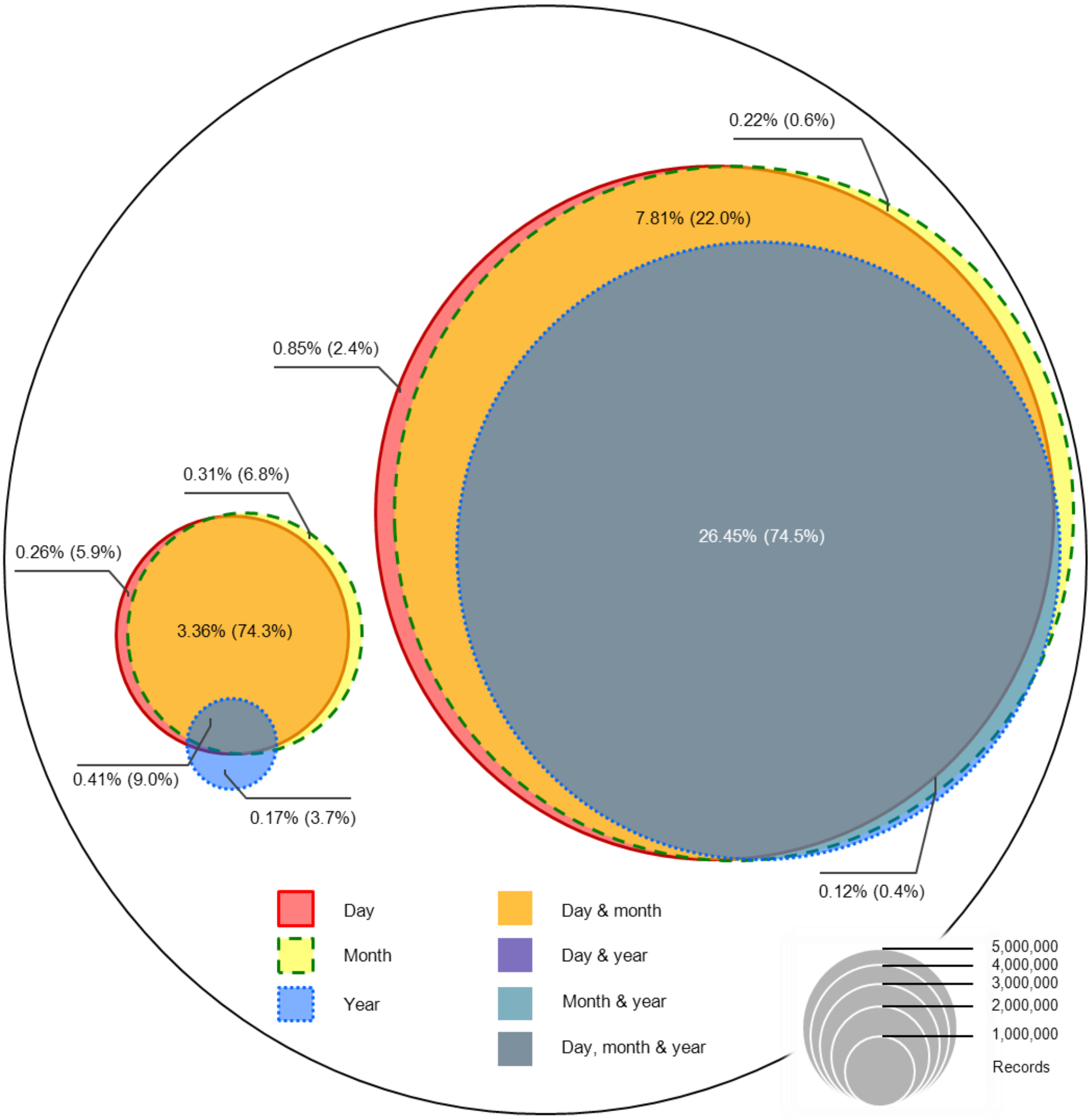
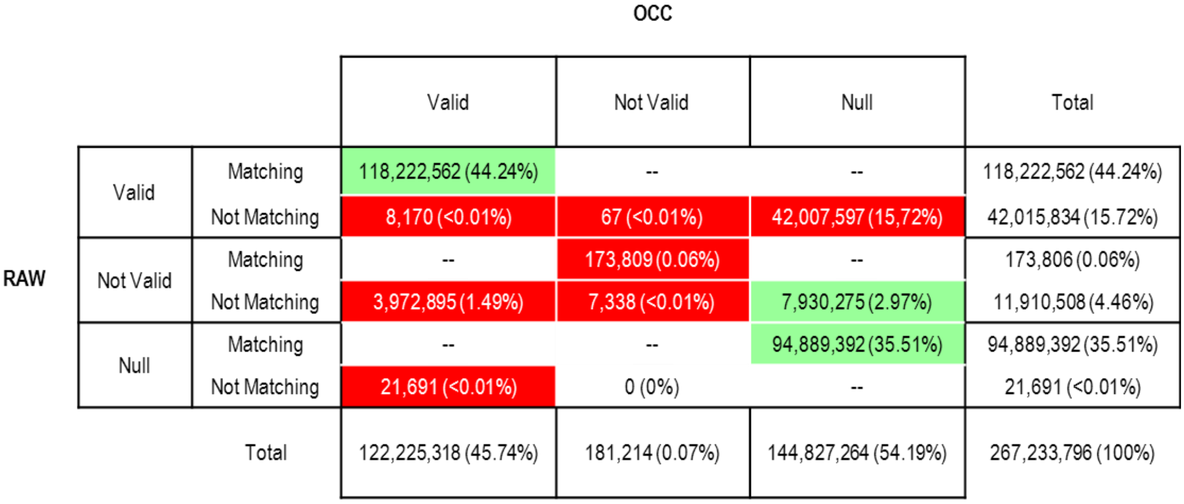


Figure 2: Causes of date invalidation (left) and nullifying (right). Proportional-area Venn diagrams (Rodgers et al., 2010). Left group: records reporting a non-null, invalid date (no date field is null); right group: records having at least one null date field. Percentages are relative to the total records in the RAW table (white circle), or to the number of records within each group (in brackets). Percentages smaller than 0.1% not shown. Areas for each group are exact, intersections are approximate (Chow & Rodgers, 2005).

## Mismatch types, volumes and sources

Nearly 83% of the records present a ‘good’ combination of date status (fig. 3), while 16% (most of what remained) were concentrated into one single case of ‘bad’ combination: nullifying OCC date when there is a valid date in RAW. From the remaining one percent of cases, almost all records were not valid RAW records that became valid in OCC. In principle, this transformation may have been either “bad”, when a valid (i.e. existing) and actual date was changed into a valid but different (unreal) date, or “good”, when a valid (existing) but not actual date became a valid, correct date.



| | | | OCC | | | |
|---|---|---|---|---|---|---|
| | | | Valid | Not Valid | Null | Total |
| RAW | Valid | Matching | 118,222,562 (44.24%) | -- | -- | 118,222,562 (44.24%) |
| | | Not Matching | 8,170 (<0.01%) | 67 (<0.01%) | 42,007,597 (15,72%) | 42,015,834 (15.72%) |
| | Not Valid | Matching | -- | 173,809 (0.06%) | -- | 173,806 (0.06%) |
| | | Not Matching | 3,972,895 (1.49%) | 7,338 (<0.01%) | 7,930,275 (2.97%) | 11,910,508 (4.46%) |
| | Null | Matching | -- | -- | 94,889,392 (35.51%) | 94,889,392 (35.51%) |
| | | Not Matching | 21,691 (<0.01%) | 0 (0%) | -- | 21,691 (<0.01%) |
| | | Total | 122,225,318 (45.74%) | 181,214 (0.07%) | 144,827,264 (54.19%) | 267,233,796 (100%) |

Figure 3: Match between dates in RAW and OCC tables. “Not valid” dates are fully-specified dates (all three fields for year, month and day filled in RAW; date field filled in OCC) that do not translate into a valid date. “Null” dates correspond to null date field in OCC and at least one null field. Red cells indicate problematic records, either because there is a mismatch significantly affecting the interpreted date, or because an invalid date in RAW made its way to OCC. Percentages are relative to the total number of records in RAW.

We searched for patterns on these transformations, and suspected that the observed differences could be attributed to the algorithm used to build the date field in the occurrence table.

Null values in the raw table resulted in null values for the date fields in the occurrence table, as designed. For example, a raw record having month as null would result in a null occurrence date, even though a filled year field may have existed in the RAW table. However, zero- or invalid (e.g., negative) numeric values for month and day in the raw table were not treated similarly. They produced a “valid”, but wrong, date in OCC. Therefore, routines to detect intrinsically invalid dates (such as validation rules: month above 12, day exceeding month’s limit, negative dates, out-of-bound years) failed, allowing for errors in dates to make their way into the data made available through the portal (the occurrences table, OCC).

When a month was supplied in the RAW table as zero rather than null, the date-building routine apparently transformed it by adding 12 to the month, and substracting one from the year. For example, a date represented in the RAW table as 20-00-1950 (dd-mm-yyyy) would become 20-12-1949. Conversely, months greater than 12 were substracted by 12 and their year added with one: 20-13-1950 would become 20-1-1951. In both cases, the “date” was apparently valid, although the year derived from the date and year fields in the OCC table would not match. There were about 3.7 million records in the OCC table affected by this problem.

Days represented as zero in the corresponding RAW table field were treated equally after month treatment, compounding the date discrepancy. The divergence could be large. For example, a date expressed in RAW as 0-0-*X* (a common way to indicate unknown month and day but known year by many publishers) had month 0 from year *X* becoming the last month (12) of the year before (*X*-1), and next, day 0 from month 12 became the last day of the month before, November 30, thus resulting into 30-11-(*X*-1).

Therefore, the origin of these issues seemed to arise from two separate but concurrent conditions: a particular way to represent missing data by data publishers (zero rather than null), and a flaw in the date validation filter at GBIF treating non-null zeroes. The flaw may have been related to a wrap-around or forcing routine for out-of-bounds or incomplete dates, either intrinsic to the database system used for the cache or implemented within the indexing code, although we cannot know the actual case (only its effects): rather than hiding (nullifying) those incomplete dates, the filter transformed them into valid, unsuspicious dates.

The values for the field 'year' in RAW and OCC had also discrepancies in 18% of records, largely arising from the same date issue but also because of other effects. Some details related to data gap analysis are provided in Gaiji et al. (this volume). About half of the issues resulted from 'year' acquiring some value in OCC when the corresponding RAW record did not have any. In many cases when no 'dateCollected' value existed in RAW, 'year' was extracted from 'dateIdentified'. Unfortunately, no further analysis was possible. It should be noted that 82% of records did match correctly.

### DISTRIBUTION OF ISSUES AMONG COMBINATION TYPES AND DATA PUBLISHERS

We attempted to characterize the extent and distribution of different types of date issues among data publishers, seeking to uncover whether the problems were widespread or focused in a few publishers contributing to the overall quality issue. Figure 4 (A) shows the series of publishers from the first to enter GBIF (left) to the latest (right). The types of issues are arranged in rows. The upper half of the plot represents the fraction of each publisher's data that is affected by each type of the date issues, while the lower half represents the fraction of total affected records across all publishers. The shade scale is in octaves (binary log). It is apparent that while many publishers contribute problematic records, these represent a small fraction of their data in most cases, typically below 1%, except for some 5% of publishers having a larger fraction of records showing issues. However, these publishers represent a relatively small fraction of the full mass of issues, except for two publishers: data publishers #139 and #172, who contributed more than 71% of problematic records. Also, most records were affected by one type of issue: invalid, non-null day and month (see also fig. 2).

We repeated this analysis for each case of "bad" date combination, focusing on the three issues affecting the largest number of records: OCC year creation from scratch, nullifying valid dates, and validating invalid dates.

Figure 4 (B) shows the distribution year value creation, split into two subcases: the creation of an invalid (top) or valid (bottom) year. More than 95% of the 13,399,281 records receiving a valid year in OCC where none was declared in RAW were attributed to just two data publishers (#169 and #111). Although this type of issue occurs in only 5% of records, these two publishers in particular are greatly affected (40% and 90% of their records, respectively).

Validation of invalid dates affected 1.49% of the total amount of records (fig. 4, C), whereas valid date nullifying affected 15.71% of the total volume (fig. 4, D; also see fig. 3). We found that one single data publisher (#139) represented 75% of this erroneous record combination, while another data publisher (#10) represented 99.5% of all cases of nullification of valid dates.

Since in most cases an individual publisher was linked to a particular type of error, selecting corrective actions should be relatively straightforward. However, many publishers were providing more than one resource. Therefore, actions could be directed at the resource or at the publisher level, depending on whether issues are tied to any specific resource, or to all resources from a data publisher.

### DATA PUBLISHER-CENTRIC PERSPECTIVE

We analysed the data contents from publishers that concentrated date issues, aiming at providing insight to suggest specific corrective policies. We focused on wrong date transformations and not-null invalid date distribution.

For 46.2 million (17.28%) records, the dataset underwent a "bad" transformation on their date information (fig. 3). Figure 5 shows the distribution of these records among data publishers. Most issues were concentrated in two publishers: Data publisher #10 represented more than 90% of the problematic records, and data publisher #139 accounted for 6.2%. Data publisher #10 contributed a number of resources, but resource #43 (representing 99.15% of the publisher's 42.2 million records) was the only one affected by this issue.

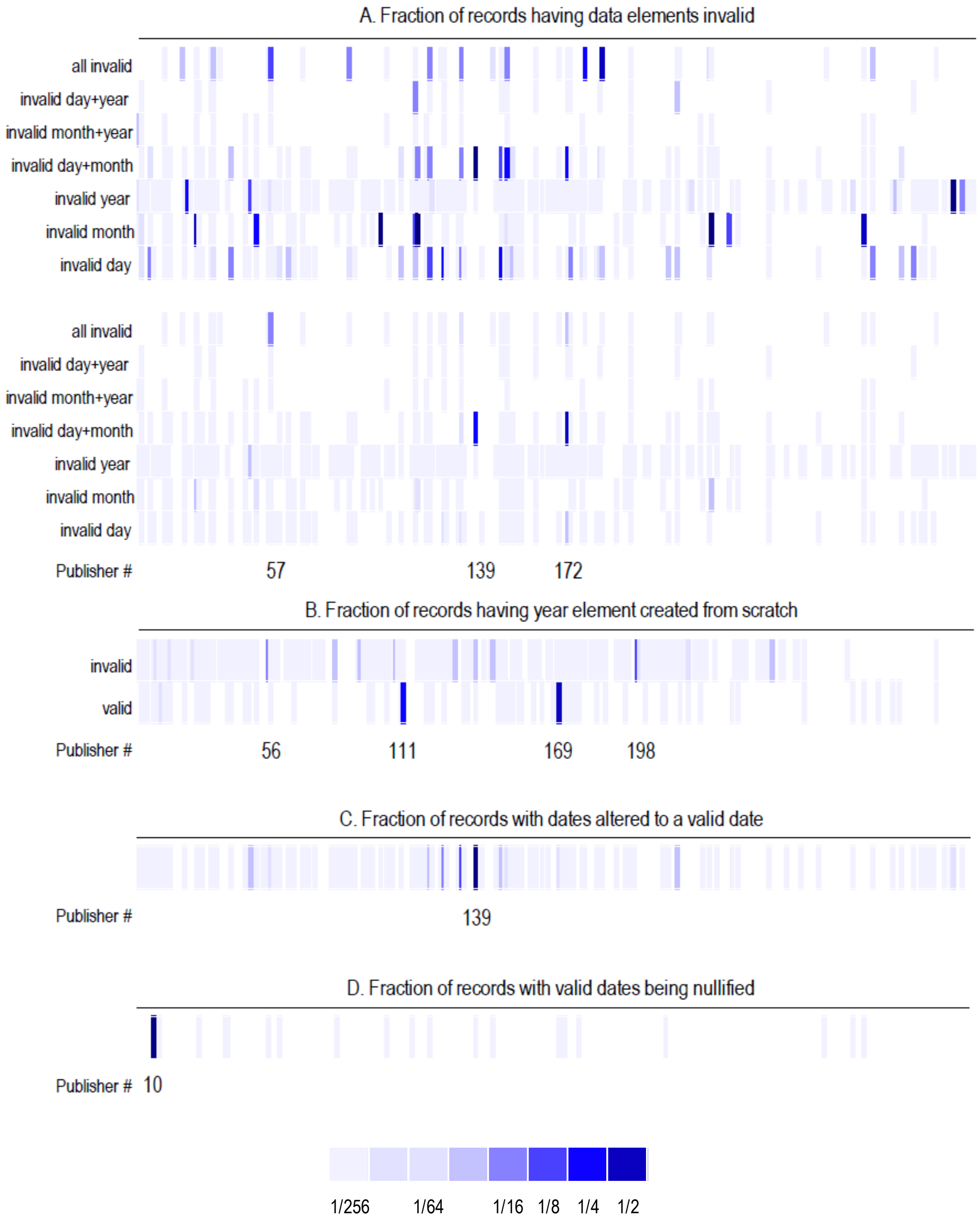


Figure 4: (A) Relative distribution of invalid, non-null dates among data publishers and error types (top: fraction of records showing issues from each publisher's total records in each category; bottom: fraction of total records in RAW that are invalid belonging to particular publishers). (B) Fraction of records in OCC from each publisher receiving a year when no year existed in date fields in RAW. (C) Fraction of records where a valid date in RAW was altered to a valid date in OCC. (D) Fraction of records where a valid date in RAW became nullified in OCC. Fractions as shades in log(2) scale ("octaves"). Series B to D are relative to the total amount of records having the issue represented in the group.

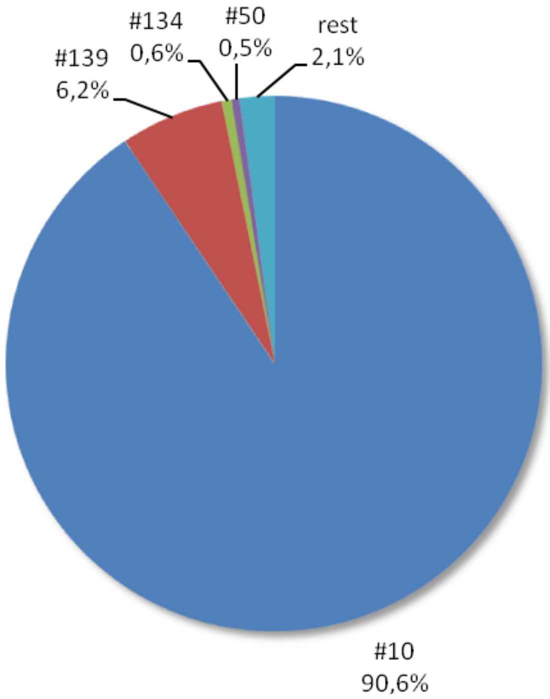


Figure 5: Distribution of records having a 'bad' date processing among data publishers.

The origin of this issue might lie on a systematic data processing problem (perhaps a data mapping problem) for this specific resource, rather than in the publisher. In the case of data publisher #139, 56.35% of its 5.1 million records had non-null zero values on day and month. Some data publishers seem to use a zero-based system for date storage (Robertson, pers. comm.), but this did not appear to be the case with this publisher as there were plenty of months specified as "12" within its resources.

On the other hand, 12 million records (4.52%) showed an invalid RAW date (table 1). Figure 4 (A) revealed the data publishers most related to this issue. The case with data publisher #139 has been discussed, but data publisher #172 is slightly different. Instead of zero or null values, this publisher stored dates in 14.56% of its 41.61 million records as empty fields, which were later interpreted as zero by the filter.

## COMPARING PROVIDED AND COLLATED DATA

Since GBIF is the main biodiversity data index, tapping on hundreds of sources, questions arise as to how many data (and how many *quality* data) can be made easily available, when these data were collected (Gaiji et al., 2013, this volume), and what can be expected in the future in terms of availability (Otegui and Ariño, work in progress). Analysis suggests that there is a bias both in data availability and data quality induced by a few data publishers. Therefore, addressing specific data publishers contributing with concentrated datasets, or showing dubious patterns, may be instrumental in dramatically improving the overall data quality at least regarding the time data in the dataset.

To take a first glance at the effects these concentrated issues would represent, we compared the number of records dated each year being published through the portal after processing (OCC table: Gaiji et al., 2013, this volume), to information in the RAW table as provided by the publishers, with and without data contributed by data publishers concentrating most date issues.

Figure 6 shows a plot of data records in the November 2010 version of the index for occurrences in the XX and XXI century. Data from 2010 onwards were not used, as we believed these were too recent to have had the same opportunity to being entered in the databases as previous years' data. The thick blue line uses data from the year field in the RAW table that is fed from publishers, while the thick red line uses the year data collected from the occurrence date in the OCC table that is published. We removed data from the data publishers concentrating most date issues, *i.e.*, #10 and #172, and plotted the filtered data (thin blue and red lines). Both lines in the filtered series (from RAW and from OCC) appear much closer than lines in the unfiltered series, and although differences exist (as can be expected from the existence of data publishers not giving exact dates), the patterns generally match, while differences are readily apparent among the unfiltered series.

The differences between both sources for date are a good indicator of a gap in quality: the relatively large amount of data coming from data publishers having date issues that prevent their transformation in valid dates, as described by the differences between the data series from the raw and the occurrences tables, point to the need for data on dates to be improved throughout the data publishers, but especially by the data publishers currently sharing the biggest amount of data through the GBIF network.

## CONCLUSIONS

We are facing a significant issue. Almost 20% of the temporal information of the openly available biodiversity data contains some type of inconsistency that reduces the quality of the main body of data. This loss of information is two-pronged: first, a lower offering of valid date information (valid date nullifying) and second, a higher volume of apparently valid although unreliable dates. These problems seem to require the convergence of two separate phenomena: the supplying of missing or invalid dates by the publishers (already a large data gap by itself, affecting almost 40% of records), and a flaw in the data validation filter.

Upon detection, GBIF should not alter the data (Andrea Hahn, pers. comm.), although it can and does flag it. Publishers and users have access to flags marking issues with records through the GBIF portal, raised at indexing time. Besides allowing the passing through of evidently invalid dates into the searchable index, the date parsing routines apparently can modify certain invalid dates so that they appear “right” and therefore escape further scrutiny by validation routines. This silent modification of records might become less problematic if the filter were more conservative, for example if all invalid values were nullified before the date to be presented to the portal is built, at the cost of losing part of the information in the date (for example, the correct sections of the triplet day-month-year). The implementation of certain validation routines previous to data publishing would allow these issues to be detected before they are made available online.

### *Recommendations for data publishers*

- Ensure that the datasets to be shared conform to the existing standard schemes, by making use of data validation and schema validation tools such as, for example, DarwinTest (Ortega-Maqueda & Pando, 2008).
- Actively consult and retrieve the event logs for the shared datasets from the portal (data.gbif.org).
- Take steps to correct issues annotated by GBIF’s parsing routines at indexing time (flag “extractTemporalParseIssue”, available in the portal’s log console), looking also for systematic sources of errors in the datasets.
- Review in-house date input policies and, if lacking or lax, enforce date validation rules using null values for missing elements of dates, database permitting.
- Set a homogeneous criterion for date mapping, using separate, null-enabled elements at the source for year, month and date.
- Document treatment of missing dates or elements thereof within the datasets’ metadata.

### *Recommendations for GBIF Secretariat and GBIF network*

- Make the date filter stricter, nullifying invalid, noninterpretable fields before reconstructing OCC’s date field. This should reduce retrieval of records with rogue dates when querying or limiting by date through the portal.
- Flag all records incurring into dubious null date components (i.e. RAW zero, empty or null) with an “incomplete” mark.
- Allow only retrieval by date on the non-null components of year or month if the record is flagged as incomplete.

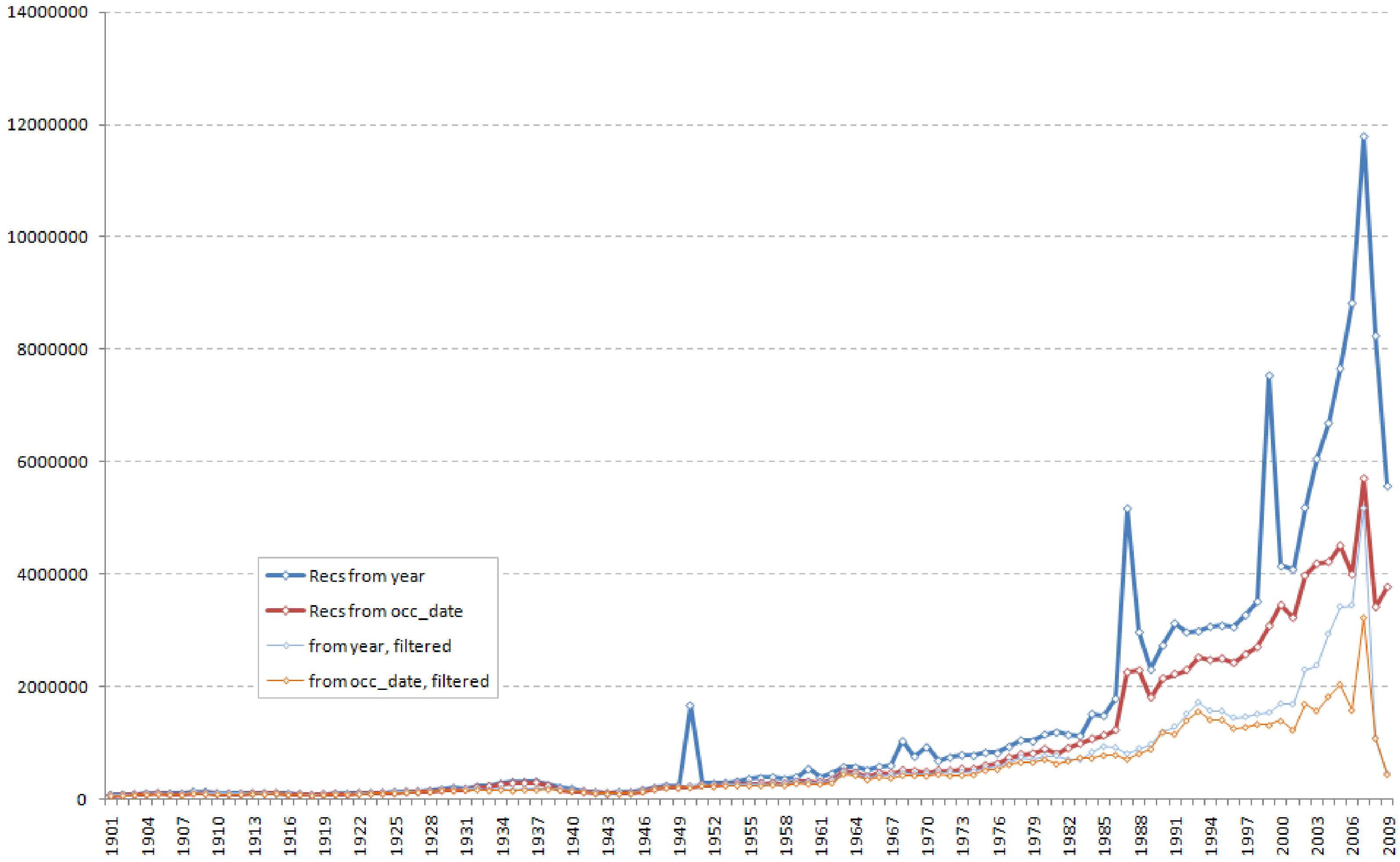


Figure 6: Number of records in the November 2010 version of the GBIF index for occurrences in the XX and XXI century. Thick blue line uses data from the year field in RAW. Thick red line uses the year collected from the occurrence_date in the OCC table. Thin lines are the same series, respectively, once removed from records contributed by data publishers #10 and #139. Data from 2010 onwards not used.

- Develop a tool to examine the level of date inconsistencies of each data set and data publisher, issuing a "date consistency index" that may indicate some systematic issue and signal for further action on the resource.

- Actively encourage data publishers to ensure the date quality of their data by providing timely feedback based on the date consistency readings, always before the next index rollover.

### *Recommendations for users*

- Review the flags of retrieved records for potential issues with dates detected at indexing time.

- Assess the fitness-for-use of the data (Hill et al., 2010) before any use for research, management or policy making.

- Use GBIF index as designed: as a discovery tool for the data being made available by the data publishers, rather than an authoritative source for end data which is not.

## ACKNOWLEDGEMENTS

Authors would like to thank Andrea Hahn and Tim Robertson for very valuable comments on the harvesting procedure at the GBIF and the current developments and for critical review of the manuscript, the staff at the Global Biodiversity Information Facility Secretariat for their support and for giving us access to the full index, and Mark Graham and one anonymous reviewer for useful remarks to the manuscript.

J.O. has received financial support for this work from the "Friends of the University of Navarra" association.

## REFERENCES

Boakes E.H., McGowan P.J.K., Fuller R.A., Chang-qing D., Clark N.E., O'Connor K., Mace G.M. (2010). Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biology 8(6): e1000385

Chow S., Rodgers P. (2005). Constructing area-proportional Venn and Euler diagrams with three circles. In *Euler Diagrams Workshop 2005:* 1-4. Accesible online at http://www.cs.kent.ac.uk/pubs/2005/2354/.

Chapman A.D. (2004). Environmental Data Quality - A Discussion paper. CRIA Report No. 5, Appendix H.

Darwin Core Task Group. (2009). Darwin Core. Available online at: http://rs.tdwg.org/dwc/

Gaiji S., Chavan V., Ariño A.H., Otegui J., Robles E. and King N. (2013). Content assessment of the primary biodiversity data published through GBIF network: Status, Challenges and Potentials. Biodiversity Informatics, 8.

GBIF (2008). GBIF Work Programme 2009-2010. Copenhagen: Global Biodiversity Information Facility. 59pp. Accessible at http://www2.gbif.org/WP2009-10.pdf.

Hill A.W., Otegui J., Ariño A.H., and Guralnick R.P. (2010). GBIF position paper on future directions and recommendations for enhancing fitness-for-use across the GBIF network, version 1.0, Copenhagen: Global Biodiversity Information Facility, 25 pp. ISBN: 87-92020-11-9. Accessible online at http://www2.gbif.org/GPP-Final.pdf

Ortega-Maqueda I., Pando F. (2008). DARWIN_TEST: Una aplicación para la validación y el chequeo de los datos en formato Darwincorev2 o Darwincore1.4. Spanish GBIF Node, GBIF.ES. CSIC. Ministry of Science and Education, Spain. Available at: http://www.gbif.es/Darwin_test/Darwin_test

Rodgers P., Flower J., Stapleton G., Howse J. (2010). Drawing Area-Proportional Venn-3 Diagrams with Convex Polygons. *In: Diagrams 2010*, LNCS (LNAI) 6170, pages 54-68. Springer. Accesible online at http://www.cs.kent.ac.uk/pubs/2010/2989/content.pdf

TDWG. (2007). Access to Biological Collections Data. Accessible online at: http://rs.tdwg.org/abcd/2.06/rddl-2007-10-18.html

TDWG. (2010). TAPIR – TDWG Access Protocol for Information Retrieval. Accessible online at: http://www.tdwg.org/dav/subgroups/tapir/1.0/docs/tdwg_tapir_specification_2010-05-05.htm

Yesson C., Brewer P.W., Sutton T., Caithness N., Pahwa J.S., Burguess M. et al. (2007). How global is the Global Biodiversity Information Facility? PLoS ONE 11: e1124.