# Scientific Datasets: Informetric Characteristics and Social Utility Metrics for Biodiversity Data Sources

**Peter Ingwersen**

**Abstract** The contribution places biodiversity datasets in relation to other central elements of the modern scientific communication system and defines quantitative analyses of metadata of such datasets as belonging to the intersection of Scientometrics and Webometrics. The analyses show that rank distributions of social utility evidence, such as search events and retrieved and viewed dataset records over a given range of datasets follow power law characteristics. A variety of dataset usage index (DUI) metrics is exemplified and illustrated by dataset indicators from three large, medium and small US and Danish dataset providers observed over a one-year period and compared to recent developments. Metrics discussed are of absolute as well as relative nature and include popularity, social attractiveness, and usage and interest impact scores.

**Keywords** Science communication · Biodiversity datasets · Webometric analysis · Social utility; Altmetrics · Dataset usage · Usage indicators · Rank distributions · Power law

## Introduction

Scientific datasets are becoming increasingly vital to understand as a central component of the modern scientific communication process—Fig. 1. Like for academic publications indexed in traditional citation databases, such as the Web of Science, PubMed or SCOPUS, entire datasets do rarely become deleted from the database or archive. Their original records are rarely edited or erased; but datasets, in particular biodiversity datasets, may indeed be updated and grow in number of records over time or be modified or restructured. This characteristic is associated with the

Peter Ingwersen, Professor Emeritus, Royal School of Information and Library Science, University of Copenhagen, Denmark. Research areas: Interactive IR evaluation and theory; Webometrics-Scientometrics; Research evaluation. Awarded the Derek de Solla Price Medal, 2005 and the ASIST Research Award, 2003. He has initiated the CoLIS and IIIX conferences and published several highly cited research monographs and journal articles on IR and Scientometrics. Telephone: +45 32 34 15 00; E-mail: clb798@iva.ku.dk

P. Ingwersen (✉)
Royal School of Library and Information Science, University of Copenhagen
Birketinget 6, 2300 Copenhagen, DK, Denmark
e-mail: clb798@iva.ku.dk

C. Chen, R. Larsen (eds.), *Library and Information Sciences,*
DOI 10.1007/978-3-642-54812-3_8, © The Author(s) 2014

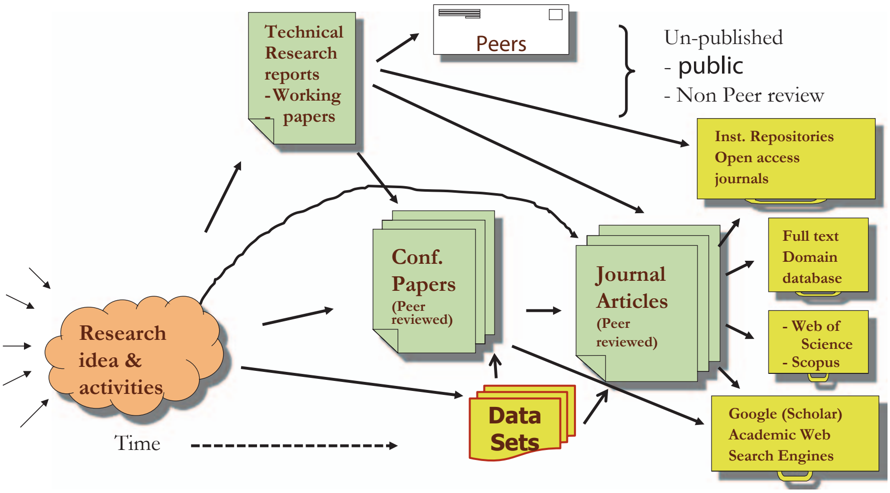

Technical Research reports
- Working papers
Peers
Un-published
- public
- Non Peer review
Inst. Repositories Open access journals
Conf. Papers (Peer reviewed)
Journal Articles (Peer reviewed)
Full text Domain database
- Web of Science
- Scopus
Research idea & activities
Data Sets
Google (Scholar) Academic Web Search Engines
Time

**Fig. 1** The scientific communication process. Revised from Ingwersen (2011)

potential for change also observed in many Web-based documents. However, unlike references given in academic publications crediting influence or direct knowledge import from other publications no common standards are available for crediting scientific datasets across the array of disciplines (Green 2009). Thus, none of the aforementioned citation-based systems explicitly take into account scientific datasets as targeted objects for use in academic work.

For biodiversity data a task force was working on this issue in order to generate recommendations for the foundation of a workable citation mechanism (Moritz et al. 2011). In addition, a set of Data Usage Index (DUI) indicators has been developed (Ingwersen and Chavan 2011). The central indicators for the development of a DUI were based on search events and dataset download instances. The DUI is intended also to provide novel insights into how scholars make use of primary biodiversity data in a variety of ways. Similar to scientometric analyses applying rank distributions, time series, impact measures and other calculations based on academic publications (Moed 2005), the social usage of primary biodiversity datasets has led to observations of their statistical characteristics as well as the development of a family of indicators and other derived significant measures. The indicators can be regarded a kind of social utility metrics which, like citations, ratings or recommendations, may be applied as impact measures in research evaluation and form supporting relevance evidence for retrieval purposes (Ingwersen and Järvelin 2005).

Initially, the presentation places the biodiversity dataset indicators within the framework of Informetrics, as a sub-section of scientometric analysis and associated with Webometrics. This is followed by examples of selected rank distribution properties of biodiversity datasets in order to observe if such distributions are similar to those observed for academic journals and articles, i.e. if they follow Bradford-like long-tail distributions. In such power-law-like cases it is expected that information management solutions similar to those used in repository management and libraries can be applied to biodiversity datasets. In addition, one may expect such statistical properties to lead to useful social utility-based research monitoring metrics. A selection of DUI indicators that are useful from this perspective, such as *Usage* and *Interest Impact* scores and relative data usage impact, will be highlighted and exemplified. The presentation ends with a brief discussion of consequences of the biodiversity dataset characteristics from the perspectives of dataset management, retrieval and evaluation.

## Biodiversity Datasets in the Informetric Framework

The scientific communication system displayed in Fig. 1 (Ingwersen 2011) contains several key components that may serve as fix points for scientometric indicator developments. Foremost they center on official research output, such as conference proceeding papers and journal articles, but also monographic publications, working papers and research reports are relevant in this respect. Patents (not shown on the Figure) signify additional particular kinds of research output, with own databases and indicator systems. With increased accessibility through the Web institutional repository publications as well as a growing body of scientific datasets of various kinds are available to researchers. In particular, datasets are used and re-used in order to carry out many different kinds of analyses, e.g. meta-analyses; benchmarking; bio topographic studies; genomics analyses, etc. Like for publications, datasets can be analyzed for their properties, for instance, with respect to volume of records, objects or topics they index and describe, and properties of authorship. Biodiversity datasets are interesting, because most are available on the Web often in a standardized database setting, but they require a lot of work to establish and this resource is only indirectly credited in the publications actually relying on biodiversity datasets. Thus the development of the set of DUI indicators analyzed below.

By being accessible on the Web one might argue that biodiversity dataset indicators based on social usage (on the web) belong to Webometrics, alternatively to the range of so-called ‘altmetrics’ indicators (Kurtz and Bollen 2010), Fig. 2. Webometric analyses imply quantitative studies of the Web, including usage of web-based resources. ‘Altmetrics’ has recently been proposed as a sub-area of Webometrics fundamentally dealing with the study of usage of social media (on the Web) such as Twitter, Facebook, blogs, and similar social networks. Typically, the actual usage population is fairly unknown in ‘altmetric’ analyses—as in many but not all webometric research areas—implying that the statistical properties are difficult to assess or control. In biodiversity dataset usage this is also the case: who is behind the searching computer is unknown to the online analyst, but the geographical area from which the search is done is known to the biodiversity dataset server. In addition, some properties are well known: the affiliation of the dataset provider; the size of the dataset in question; the topics and objects covered by the dataset.

It is thus fair to state that Informetric analyses of biodiversity datasets belong to Scientometrics, i.e. quantitative analyses of the science system(s), using Bibliometric methods, such as rank distributions, and intersected with Webometrics since the datasets are available through the Web, Fig. 2. Whether to use the notion

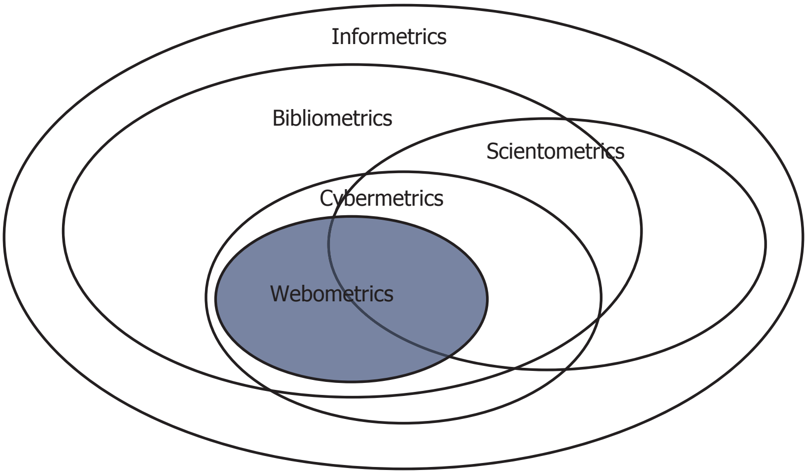


**Fig. 2** The framework of Informetrics (from Björneborn and Ingwersen 2004, p. 1217).

of 'altmetrics' or simply webometrics for the analyses made is an open question, which I as instigator of Webometrics as a study area (Thelwall et al. 2005) will let the community to decide.

## Biodiversity Dataset Characteristics

The objectives of the proposed DUI were (Ingwersen and Chavan 2011, p. 2) "[to] make the dataset usage visible, providing deserved recognition of their creators, managers, and publishers and to encourage the biodiversity dataset publishers and users to:

- Increase the volume of high quality data discovery, mobilisation and publishing;
- Further use of primary biodiversity data in scientific, conservation, and sustainable resources use purposes; and
- Improve formal citation behaviour regarding datasets in research."

In order to do so understanding of the characteristics of the datasets and their behaviour in the scientific life-cycle is central. Biodiversity datasets are presently accessible online through the Global Biodiversity Information Facility (GBIF) located in Copenhagen, Denmark. The structure and prospects of GBIF is outlined by Chavan and Ingwersen (2009). The GBIF data portal was established in 2001 (http://data.gbif.org.) and holds currently over 400 million records published in more than 10,000 datasets by almost 500 data publishers, with the largest data set containing more than 21 million records. The Data Usage Index (DUI) indicator developments were based on data usage logs of the GBIF data portal. The logs provide general usage data on kinds of access and searches via IP addresses as well as download events of datasets within the control of the GBIF data portal. As a spin-off the usage logs also provides different rank distribution characteristics, which are directly

**Table 1** Top-18 distribution of search events and number of records viewed, ranked by Search Events in the Danish Biodiversity Information Facility (DanBIF); (GBIF, December 1–31, 2009). Search density signifies no. of records per search event

| Data set | Search events | Rank by no. of rec. | Searched records | Search density |
|---|---|---|---|---|
| Danish Mycological Society, fungal records database | 1149 | 2 | 32394 | 28.2 |
| Botanical Museum, Copenhagen, Mycology Herbarium | 1035 | 3 | 22242 | 21.5 |
| Niva Bay species list, Sjalland, Denmark | 387 | 18 | 2834 | 7.3 |
| Heilmann-Clausen) | 372 | 6 | 9145 | 24.6 |
| DOF | 339 | 1 | 35758 | 105.5 |
| Galathea II, Danish Deep Sea Expedition 1950–52 | 329 | 7 | 8912 | 27.1 |
| Priest Pot species list, Cumbria, Britain | 325 | 15 | 3520 | 10.8 |
| Herbarium | 249 | 4 | 19925 | 80.0 |
| Western Palearctic migrants in continental Africa | 191 | 5 | 13655 | 71.5 |
| Botany registration database by Danish botanists | 172 | 11 | 5083 | 29.6 |
| Palaearctic | 161 | 10 | 5556 | 34.5 |
| DOF 2001–2006 | 158 | 8 | 8560 | 54.2 |
| University's Arboretum | 152 | 12 | 4714 | 31.0 |
| Marine Benthic Fauna List, Denmark | 137 | 19 | 2532 | 18.5 |
| Botanical Museum, Copenhagen, the Lichen Herbarium | 133 | 14 | 3618 | 27.2 |
| Botanical Museum, Copenhagen, type specimens | 60 | 30 | 161 | 2.7 |
| Danish Ants (Formicidae) | 56 | 13 | 3952 | 70.6 |
| Galapagos grasses and sedges | 54 | 29 | 194 | 3.6 |

accessible online for analysts through the GBIF Portal and its datasets—eventually via known dataset providers.

Table 1 demonstrates the top-rankings of a typical distribution of different datasets produced by the same dataset provider (Biodiversity data: Danish Biodiversity Information Facility, DanBIF, GBIF 2010) at a specific time period, i.e. one month, December 2009. During the selected time slot the provider was searched in total 5,704 times and the users looked at 207,622 records from the 36 available datasets, with an average search density of 36.4 records. Out of this volume the GBIF logs inform that 42,923 records were downloaded through 538 download events (average download density = 79.8 records), not shown on Table 1. Like for journal articles distributed over a publishing journal according to citations, the GBIF mobilized dataset records might be distributed over datasets according to usage (downloads) or searching.

Detailed analyses of the GBIF logs reveal that similar to articles vs. journals a Bradford distribution can be observed for searched biodiversity dataset records dispersed over datasets. A Bradford rank distribution of journals is a Gini-index

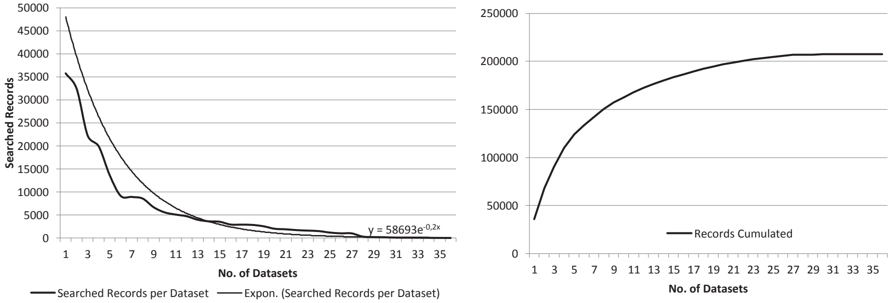


**Fig. 3** Rank distribution of 36 datasets in DanBIF according to searched records (GBIF. December 1–31, 2009). Actual distribution (*left*) and cumulated distribution (*right*)

like distribution of the power law form *a; an; an*$^2$—where *a* signifies the number of journals publishing the upper tertile of articles (≈ datasets producing the upper tertile of records, searches or downloads) and *n* a constant specific for that scientific area (Garfield 1979; Moed 2005). Although the number of datasets in the distribution is quite small (36) we may, with good will, observe an approximation to a Bradford distribution for the searched records: the first tertile (69,207 records) of the total number of records (207,622) is covered by the top-2 1/2 datasets alone (sorted by Record Number = 79,273 records). The next 6 1/2 datasets cover 74,086 records, approximating the second tertile. The remaining 27 datasets cover the last tertile. This approximates to $a = 2.5$ datasets; $a\ n = 2.5 \times 3$ (= 7 1/2 datasets) and $a\ n^2 = 2.5 \times 9$ (= 22 1/2 datasets). A Bradford distribution for a given range of datasets implies that very few datasets (2–3) cover a large portion (> 33 %) of the entire volume of records in the area covered by the range of datasets (here defined by the provider), followed by a long tail phenomenon.

In fact, the pattern shown is steeper than suggested by a standard Bradford distribution. More than 2/3 of the searched records in the DanBIF biodiversity collection (142,000 records) were covered by only 7 datasets (20 %), Fig. 3, right-hand side. From Table 1 we observe that of the top-10 datasets ranked according to search events (popularity) seven datasets were also those sets with most used records as searched and viewed by peer biodiversity researchers world-wide. The pattern can be monitored over time for consistency, see example Table 2 for the HUA provider. During the monitored month in 2009 the DOF dataset was the most used set according to *Searched Records* but ranked fifth with respect to *Searching Event* frequency, i.e. popularity. In addition the DOF dataset had the highest Search Density (105.5 records per search event).

Figure 4 displays the corresponding rank distribution of search events over the 36 DanBIF datasets during the same time slot, again providing a long tail distribution, but with two datasets standing out as most searched (popular) datasets. Cumulated they constitute 38 % of all events taking place during the period (2184 search events of a total of 5704 events), Fig. 4, right-hand side.

Data can be extracted from other elements of the GBIF data portal logs in order to generate rank distributions, e.g. associated with specific species or of frequent

**Table 2** Dataset indicator examples. Record No. as per December 31, 2009

| Indicator | Formula | OBIS Dec09 | DanBIF-09b | HUA-09a | HUA-09b | DK 2009b |
|---|---|---|---|---|---|---|
| *Searched Records* | $s(u)$ | 2,092,927 | 5,682,095 | 2,299,133 | 7,328,160 | 13,010,255 |
| *Download Freq.* | $d(u)$ | 555,835 | 854,761 | 809.468 | 717,102 | 1,571,863 |
| *Record Number* | $r(u)$ | 11,140,298 | 4,995,544 | 259,077 | 259,077 | 4,836,771 |
| *Search Events* | $S(u)$ | 42,860 | 249,214 | 126,449 | 198,910 | 448,124 |
| *Download Events* | $D(u)$ | 601 | 4,486 | 2,059 | 1,760 | 6,246 |
| *Dataset Number* | $N(u)$ | 180 | 38 | 2 | 2 | 40 |
| *Datasets used* | $n(u)$ | 171 | 36 | 2 | 2 | 38 |
| *Search Density* | $s(u)/S(u)$ | 48.83 | 22.80 | 18.18 | 36.84 | 29.03 |
| *Download Density* | $d(u)/D(u)$ | 924.85 | 190.54 | 393.14 | 407.44 | 251.66 |
| *Usage Impact* | $d(u)/r(u)$ | *0.05* | *0.17* | *3.12* | *2.77* | *0.32* |
| *Interest Impact* | $s(u)/r(u)$ | *0.19* | *1.14* | *8.87* | *28.29* | *2.69* |
| *Usage Ratio* | $d(u)/s(u)$ | 0.27 | 0.15 | 0.35 | 0.10 | 0.12 |
| *Usage Balance* | $D(u)/S(u)$ | 0.014 | 0.018 | 0.009 | 0.009 | 0.014 |

The Herbarium of University of Aarhus (HUA), covers two periods: Jan-June (a) and July–Dec. (b), 2009; the Danish Biodiversity Information Facility (DanBIF) July–Dec., 2009 (b); and Ocean Biogeographic Information System (OBIS) analyzed December 1–31, 2009 (GBIF Portal 2010; Ingwersen and Chavan 2011)

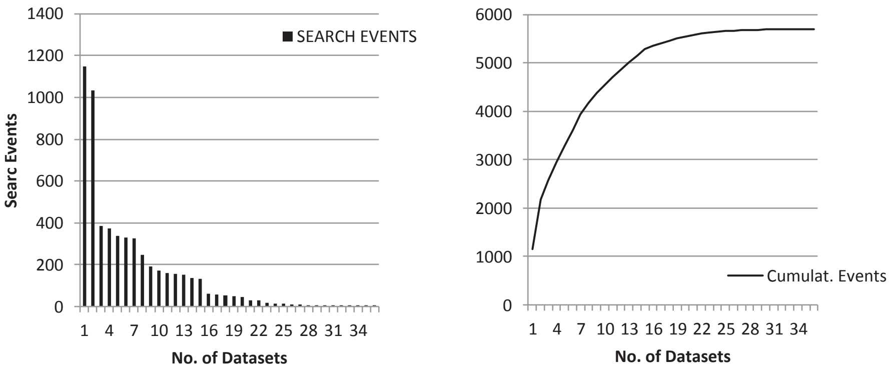
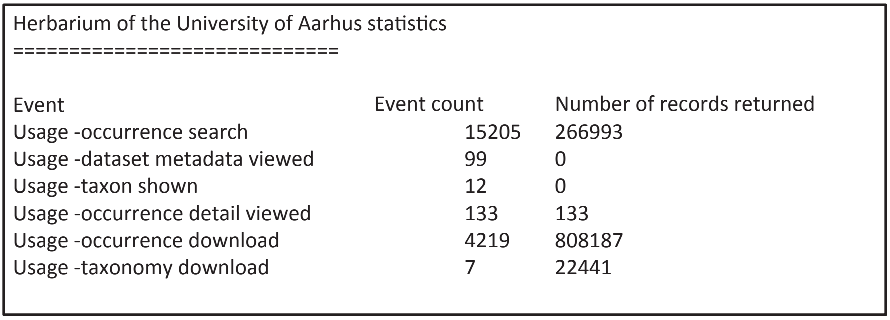


**Fig. 4** Rank distribution of search events across the 36 datasets in DanBIF (GBIF; December 1–31, 2009). Actual distribution (*left*) and cumulated distribution (*right*)


```
Herbarium of the University of Aarhus statistics
============================

Event                                  Event count    Number of records returned
Usage -occurrence search                    15205     266993
Usage -dataset metadata viewed              99        0
Usage -taxon shown                          12        0
Usage -occurrence detail viewed             133       133
Usage -occurrence download                  4219      808187
Usage -taxonomy download                    7         22441
```

**Fig. 5** Extract of GBIF Event log file downloaded covering February-August 2013. (Biodiversity data: Herbarium Database Aarhus University HAU, GBIF 2013)

visits gaining access into specific dataset providers or datasets, via IP addresses. These latter distributions rank the top players that *import knowledge* in specific areas or from particular providers, datasets or species/taxa. Only the GBIF server staff is able to extract such data whilst the shown distributions are publicly available online. As part of its architecture the GBIF data portal supplies up-to-date lists of datasets as well as of dataset publishers, sorted alphabetically and detailing dataset name, Record Number and an entry to the dataset event log. The lists and structured event logs per dataset and provider can be downloaded easily (Fig. 5) and eventually re-ranked or manipulated statistically offline.

A recent online analysis of the GBIF event log demonstrates that, for instance, the Danish Mycological Society (Row 1, Table 1) at present holds 81,000 records, and during the month December 1–31, 2012 the dataset was searched 250,001 times retrieving and viewing 5,234,732 records with a search density of 20.9 records per event (Biodiversity data: Danish Mycological Society, GBIF 2013). These figures illustrate the dramatic increase of the usage of the GBIF portal over a period of one month during three years, see Table 1 for comparison.

## Dataset Usage Index Indicators

In (Ingwersen and Chavan 2011) the range of DUI indicators is defined, exemplified and discussed. They are based on the extracts of data from the GBIF event logs for datasets and are constructed according to common scientometric standards for research evaluation indicators (Moed 2005). Below we point to the most prominent indicators and discuss briefly their potentials, since they are characteristic for biodiversity datasets that are publicly available, searched and downloaded. Table 2 demonstrates 13 of the indicators, exemplified by dataset properties from three different dataset providers: The large network-like US-based Ocean Biogeographic Information System (OBIS) and the two Danish providers DanBIF and Herbarium of University of Aarhus (HUA).

The Number of Datasets produced by a publisher (N(u)) at a given point in time may characterize the publisher into small ($N<10$), Medium ($10<N<100$), Large ($100<N<300$) and ultra-large ($N>300$). The reason behind this classification is that it is meaningless to compare between providers of quite different sizes. Like for citation impact small (often specialized) universities should not be compared to large universal universities. DanBIF is thus regarded a medium-size provider while HUA is seen as a small dataset producer. Table 2 provides an overall view of their characteristics (Ingwersen and Chavan 2011, p. 7).

For the large-scale US provider OBIS the analysis window is one month against 6 month for the two other providers. Comparisons should hence not be carried out them in between. The *Usage Ratio* signifies the number of records downloaded over number of records searched during the same period. The higher the ratio the more searched records are also subsequently downloaded and imply a kind of *social attractiveness* of the datasets in question.

The table shows that regardless of length of analysis window the numbers of *Searched Records* and *Download Frequency* were quite substantial in 2009, supporting the conception of a DUI. *Download Events* were very low compared to the number of *Search Events* across all three publishers and periods. Three years later the GBIF portal seems well established in the mind of the global research community, Fig. 5, with a *Download Event* score during the six-month period February–August 2013 for HUA raised to 4,226 events and with a *Download Density* reaching 196.6 against 15,205 *Search Events* with a corresponding much lower *Density* of 17.6 (Biodiversity data: Herbarium of University of Aarhus 2013).

The *Usage Balance* between *Download* and *Search Events* was quite low in 2009: only approx. 1–2 % of the search events lead to direct downloading for the providers; for HUA less than 1 %. In 2013, for one HUA dataset, the *Usage Balance* reaches 28 %, implying that for each 4 search events there is one pure download event taking place, signifying that searchers seem more familiar with the dataset contents and do not require constantly to search and investigate the set prior to actual usage. This coincides with the *Usage Ratio*, or social attractiveness score, which for HUA during the six month in 2013 reaches 3.1 signifying that more than three times the searched records are actually downloaded.

According to Ingwersen and Chavan (2011, p. 7) the *Interest* and *Usage Impact* factors inform about the average number of times each record stored by a dataset publisher has been searched or actively downloaded. In both metrics a value greater than 1.0 implies that in principle all the dataset records on average have been searched or downloaded at least once during the analysis period. The two time slots, Table 2 (2009a, b), may illustrate the developments for a dataset provider like HUA during the entire year 2009, i.e., showing a slight decrease in *Usage Impact* (from 3.1 to 2.8) and a strong increase in *Interest Impact* (from 8.9 to 28.3). In contrast, during the recent six-month period in 2013 HUA's *Usage* and *Interest Impact* values are 7.4 and 2.4 respectively[1]. The *Usage Impact* has increased substantially while the *Interest Impact* has noticeably dropped. This is due to a strong increase in downloads and much less searching and viewing activity during the later period, in accordance with the *Usage Balance* and *Ratio* scores.

Aside from the DUI indicators, Table 2, the event logs may in addition produce data on the most popular objects, i.e., the species in the dataset that are most searched and viewed during the selected analysis period. Such data constitutes the *usage profile* for a particular dataset and changes can be monitored over time.

These absolute DUI metrics can be turned into relative indicators, e.g. by relating single datasets to their provider's cumulated properties or associating several providers to the national aggregation for particular indicators. The HUA *Usage Impact Factor* for 2009b relative to Denmark (U-IF/DK) is thus 2.77/0.32 = 8.65. The corresponding U-IF (DanBIF) is 0.53. Examples of relative DUI indicators and all formulas are shown in (Ingwersen and Chavan 2011).

## Concluding Remarks

The presentation demonstrates the feasibility of establishing a framework for academic crediting of dataset production, searching and usage. The Dataset Usage Index signifies a step forward towards such a dataset management framework. The reason that the DUI is appropriate lies in the rank distribution properties which, among other characteristics, follow the pattern of power laws in proximity of Bradford distributions. Further, the distributions make it feasible to point to the most popular or socially attractive datasets, providers or species, monitored over time, and to apply such evidence in dataset management decisions as well as for retrieval purposes. The latter perspective reaches into types of recommendation systems commonly applied to other kinds of social media (Bogers and van den Bosch 2011). Because of their usage dimension biodiversity datasets, as well as other scientific datasets may be seen as particular kinds of cooperative filtering information systems.

In addition, a range of absolute as well as relative usage indicators has been defined and exemplified. Biodiversity datasets and their records seem to display some similar characteristics as journals and articles published in such journals. It is thus

[1] The number of records in the one HUA dataset available in August 2013 is 111,525.

very likely that information management traits that have been found appropriate for academic journals and journal articles in repositories and libraries are equally useful for biodiversity and other scientific datasets. Similarly, a DUI is likely to serve as a convenient complement to traditional citation-based research monitoring, in particular with respect to institutional evaluations since the biodiversity datasets constitute a substantial workload otherwise not made visible in traditional research monitoring schemes.

## References

Biodiversity occurrence data published by: Danish Biodiversity Information Facility (Accessed through GBIF Data Portal, data.gbif.org, 2010-01-01)

Biodiversity occurrence data published by: Danish Mycological Society (Accessed through GBIF Portal, data.gbif.org, 2013-08-16)

Biodiversity occurrence data published by: Herbarium of University of Aarhus (Accessed through GBIF Portal, data.gbif.org, 2010-01-01; 2013-08-16)

Biodiversity occurrence data published by: Ocean Biogeographic Information System (OBIS) (Accessed through GBIF Portal, data.gbif.org, 2010-01-01)

Björneborn L, Ingwersen P (2004) Toward a basic framework for webometrics. J Am Soc Inf Sci Technol 55(14):1216–1227

Bogers T, van den Bosch A (2011) Fusing recommendations for social bookmarking websites. Int J Electron Commer 15(3):31–72

Chavan VS, Ingwersen P (2009) Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. BMC Bioinformatics 10(Suppl 14):11 s

Garfield E (1979) Bradford's law and related statistical patterns. current comments; essays of an information scientist 1979–1980, May 7 1979. pp 5–11

GBIF. http://data.gbif.org/

Green T (2009) We need publishing standards for datasets and data tables. White paper OECD Publishing; 2009, 9–11. doi:10.1787/603233448430

Ingwersen P, Chavan V (2011) Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. 15 Dec. 2011. BMC Bioinformatics 12(Suppl 15):S3, s. S3. 10 s

Ingwersen P, Järvelin K (2005) The turn: integration of information seeking and retrieval in context. Springer

Kurtz M, Bollen J (2010) Usage bibliometrics. Ann Rev Inf Sci Technol 44:3–64

Moed HF (2005) Citation analysis in research evaluation. Springer, Dordrecht

Moritz T, Krishnan S, Roberts D, Ingwersen P, Agosti D, Penev Y, Cockerill M, Chavan V (2011) Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing Framework Task Group. 15 Dec. 2011. BMC Bioinformatics 12(Suppl 15):S1, 10 s

Thelwall M, Vaughan L, Bjorneborn L (2005) Webometrics. Ann Rev Inf Sci Technol 39:81–135