KOSMOS
PROBLEMY NAUK BIOLOGICZNYCH

Tom 70 **2021**
Numer 2 **(331)**
Strony **139–145**

Polskie Towarzystwo Przyrodników im. Kopernika

PIOTR TYKARSKI

Institute of Functional Biology and Ecology
Faculty of Biology, University of Warsaw
Biological and Chemical Research Centre
101 Żwirki i Wigury Str., 02-089 Warsaw
E-mail: p.tykarski@uw.edu.pl

NATURAL DATA RESOURCES OF POLISH SCIENTIFIC INSTITUTIONS – VARIETY, HISTORY, IMPORTANCE – INTRODUCTION*

The present issue of KOSMOS is devoted to the Polish resources of information on biological diversity, in particular, to source materials such as collections of specimens. The subject is discussed in articles written by custodians of natural collections and authors related to the collection and analysis of natural data, representing institutions collaborating within the Polish Biodiversity Information Network (KSIB)[1] within the project of the Digital Poland Operational Programme "Integration and mobilization of data on biotic diversity of Eukaryota in resources of Polish scientific institutions" (IMBIO)[2].

We live in the age of digitization. With rapid development of computers, digitization of data was initiated, gradually encompassing ever wider areas of human activity. Appearance of the Internet and rapid development of digital communication have further intensified this process. This trend also included biological data, with the most intensively developing fields utilizing modern computational technologies, together with progress in biochemistry and genetic research, led to a specific appropriation of the term „bioinformatics", narrowed down to the area of molecular biology.

The data resources of the organismal biology were digitized, too, albeit, due to their abundance and variety, the process occurred at different speeds and did not bring spectacular achievements in its initial phase, comparable to protein spatial modelling or genome mapping.

The breakthrough moment was the launch of the Global Biodiversity Information Facility (GBIF)[3] system, which revealed the potential hidden in so far scattered databases of the organismal level. The ability to combine data from different sources, regardless of their original type and format, has proven unprecedented – from 18th century collections of specimens, through traditional scientific publications, to systems of data collection, supplied with data from mobile phone applications (Fig. 1). The achievements in this area, its intensive development and data growth were so high that it gained the name of „biodiversity informatics", with its own technologies, tools and standards.

Nevertheless, modern technologies are only means, and the usefulness and value of a system depends on the scientific value and quality of the information. From this point of view, the most valuable assets are data coming from the oldest type of sources, i.e., collections of specimens. This is an obvious fact for specialists on the subject, while for those who are keen on development of technology and perceive the traditional "museum science" as a relic – a great discovery. This is due to the fact that it is the modern techniques, such as DNA analyses, 3D imaging, and mass data processing, that have showed the huge value of the traditional specimen collections.

[1]www.ksib.pl
[2]imbio.uw.edu.pl
[3]www.gbif.org

*This publication was created due to financial support of the project POPC.02.03.01-00-0081/19 „Integration and mobilization of data on biotic diversity of Eukaryota in resources of Polish scientific institutions „ (IMBIO).

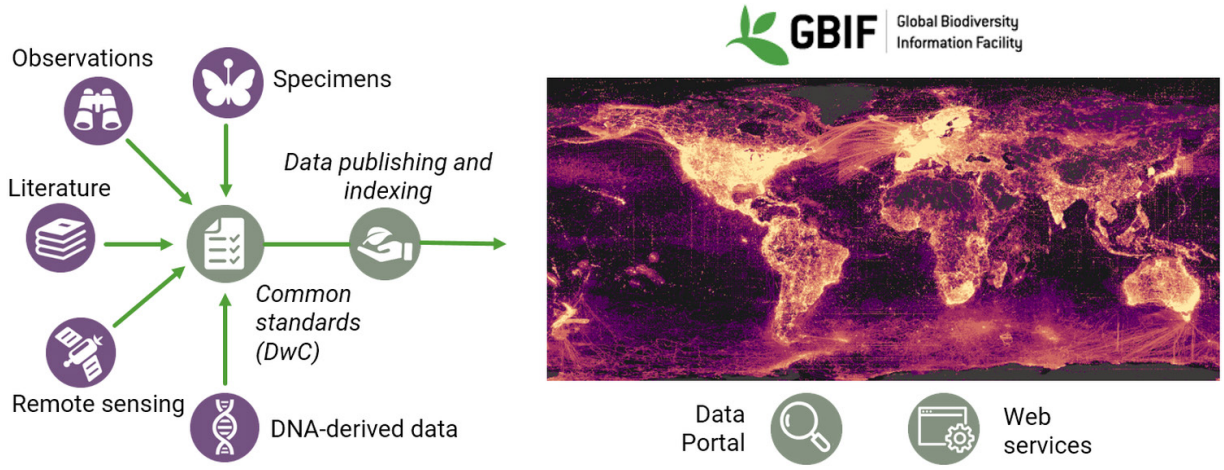

Fig. 1. Diagram of the current data flow in the GBIF.

Apart of traditional data sources – specimens and observations – the system receives ever more data from automated processing of literature, initial remote sensing data, and an increasing amount of data from genome research. Common standards of data exchange (mainly Darwin Core) allow to create an index and to publish all data sets. The resources of the network are available by the GBIF portal, and the network services offer their automatic use by external applications. The map shows data density (the number of distribution records) available in GBIF – the lighter shade, the more data (as of July 2021). *The illustration contains elements of an unpublished slide from a presentation of the GBIF Secretariat (with their consent).*

In technical terms, for an information unit (an occurrence record) in a simple database, the difference between an observation and a specimen consists merely in a different value of the "record type" field (specimen/observation). The remaining elements may be identical: both a specimen and an observation document occurrence of the species in time and space, so identical methods of describing these elements can be used, so that both types of records can be processed together in many applications.

The main difference pertains to the evidence material. For a specimen, it is a preserved biological material, i.e., a specimen or its fragment, while, for an observation, at best a photograph or another type of record of an individual's properties (video sequence, sound, etc.). In most cases, the observation is limited to the mere registration of the organism's occurrence. The credibility of such a data record depends only on the qualifications the registrar, particularly the accuracy of taxonomical identification. Correct identification is a key attribute of this information, without which the remaining details cease to be meaningful. In the case of a specimen – an individual, its fragment or product (e.g. leaf mine, gall, etc.), the existing biological material allows taxonomic identification regardless of the registration of the organism, and it offers much more possibilities for analysis than any form of recording its features. A person collecting research material may identify the specimen incorrectly, or may not identify it at all. In the case of museum materials, it is possible to permanently verify their taxonomic identification, even hundreds of years since their acquisition.

Specimens in collections are therefore an illustration, or basically a preserved fragment of the real state of the biosphere, existing in a given place and time. Properly preserved, they provide material for an ever longer list of different types of analyses of physical and chemical properties, and, consequently, an increasingly wider range of applications, limited only by the current state of science (BAKKER et al. 2020). DNA sequences preserved in specimens may be used in taxonomic analyses along with the classic morphology-based methods, opening the way to phylogenetic analyses, evolutionary studies, population and conservation genetics (WANDELER et al. 2007). Classic morphological and anatomic analyses are supplemented by non-invasion technics, such as microtomography (WIPFLER et al. 2016) or non-destructive methods of genetic analyses (SANTOS et al. 2018).

There is a dramatic difference between an observation and a specimen in terms of how much information is stored in a digital record. Even a very extensive record of all the attributes of a simple observation takes up little space compared to specimen data, for which there is no upper limit to

the data space they can occupy. Each new research technique adds new requirements in this regard, while the existing ones being improved. The simplest example is digital photography. A size of an image is a function of resolution; the subsequent generations of photographic equipment bring new features, which usually increase the requirements for the space on the hard drive or any other media. Photographic documentation of specimen collections performed at the beginning of the 21st century is a small fraction of the volume it would occupy now.

Any new research technique that uses specimens as source material requires that the results of analysis be stored, often involving files of large volumes, like in the case of 3D images. Thus each specimen can be accompanied by a set of related data sets, that are digital representations of its properties. This perception of information led to a concept of extended specimen (WEBSTER 2017, HEDRICK et al. 2020, MILLER et al. 2020), and, consequently – a digital specimen – mapping the specimen in the digital space as an open set of all data related to a physical object: from the place and date of its discovery to images and DNA sequences. In addition, meeting the FAIR[4] rules, allowing full use of the opportunities provided by the internet (LANNOM et al. 2019).

On the other hand, it is not so obvious that a digital image of a specimen may survive longer and eventually become the only record of its original material. Physical objects, especially fragile ones, such as insect bodies or soft plant tissues, undergo slow degradation. It is inevitable, one can only strive to inhibit this process by appropriate conservatory procedures. However, there is a still higher threat, i.e., the constant risk of destruction or damage from inadequate care, wrong protection means, natural disasters or hostilities. Many valuable collections have been lost due to mistakes or negligence, which finally led to the tragedy. In Poland, the most severe losses, apart from the war ones, were caused by the fire of the State Zoological Museum in Warsaw in 1935, probably due to an accident (DASZKIEWICZ at al. 2016). Among the known cases of this type around the world in recent years, the most famous was the fire of the Museum of Natural History in Rio de Janeiro, which consumed the largest collections in South America in the field of ethnography and natural history (GRESHKO 2018, MOON 2018). Unfortunately, it was just an element of a larger set of similar disasters in Brazil (MEGA 2020). However, that event shows a highly important aspect in the context of the subject matter presented in this issue of KOSMOS. A small part of the lost specimens had been digitized earlier, and their photographic images (np. FACHIN et al. 2016) and entries in databases[5] have been preserved. This is the most direct and bitter proof that digitization can be useful, even vital, and the more comprehensive and detailed it is, the more will remain when the original specimen is lost.

The scientific value of biological data related to natural collections is probably their most important aspect (at least from the naturalist's point of view), but not the only aspect. The specimens are also of cultural and historical importance, analogous to the other types of museum artifact collections. They are at the same time a heritage of the past, a legacy of passed generations, national asset and a testimony of astounding life stories of particular people who gathered them. The history of Szymon Tenenbaum's collection, presented in the biographical article in this issue of KOSMOS, is a perfect illustration of this. The national natural data resources are not sufficiently valued in the public awareness. This is due to many reasons, clearly explained in the presented articles: dramatic history of Polish scientific centers and tragic fate of many employees, since the period of the Poland's partitions, through war losses, to the times of communist rules and "real socialism" in the Polish People's Republic, when no proper conditions were provided for the development and use of scientific natural collections for research. The decades-long underfunding of units, which had to struggle to find funds for their maintenance, and negligence of the education system (which also occurred in other countries), led to a situation where fields connected with natural history were underestimated and rather unknown, even among scientists or people and institutions deciding on the scientific policy.

It is noteworthy that the collections described in this issue of KOSMOS have survived until present time owing to commitment, sometimes heroic, of individual people. They were not hindered by lack of planned, long-term strategy of the Polish state; and looking from the perspective of the history of the last 200 years, scientific collections of natural history were created and developed even in opposition to the (usually foreign) authorities. Certainly, this would not be possible without the organizational support of institutions, whose management, even in

[4]abbreviation of Findable, Accessible, Interoperable, Reusable - www.go-fair.org/fair-principles

[5]www.gbif.org/dataset/search?hosting_org=4205110f--3f0f-40d8-bd0f-2fa71bc827b5

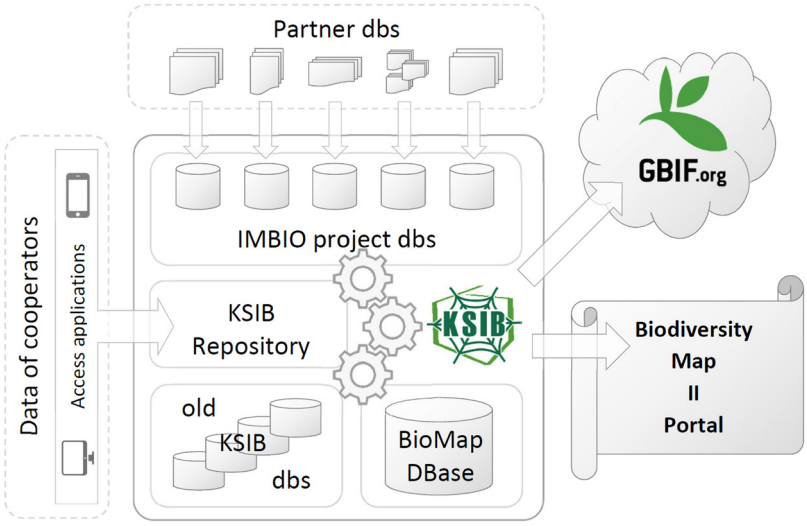

Fig. 2. Diagram of the planned KSIB IT system, integrating existing bases, resources of the Biodiversity Map (Baza BioMap) project, data from digitalization of resources of the IMBIO project partners, and the remaining resources within the constructed KSIB Repository. Introduction of data will be possible by applications (website, mobile applications). A new version of the Biodiversity Map portal will be used for data presentation. Externally, the data will be presented through the GBIF infrastructure.

most difficult periods, ensured at least the minimal conditions necessary for physical survival of these resources.

From the perspective of 20 years of biodiversity informatics development, it can be said that only recently has it started to make up for the state-wide delays in technological development and digitization of natural data resources. The first Polish databases collaborating with the GBIF[6] network appeared a in 2004[7], and since then the data resources in electronic form of institutions associated with the KSIB grew dynamically until 2008, then these activities became limited due to difficulties in obtaining long-term financial support. The effect of the first period of intensive growth was over 1.5 million occurrence records, available in 90 databases belonging to 16 institutions, of which a quarter was connected with specimen collections, and the rest based on observation data. The only way to interact with these resources was (and mostly still is) the GBIF portal.

Naturally, the works connected with the GBIF constituted only a part of activities of national centers in the field of creating natural databases and sharing their resources. This was done by member institutions of KSIB and other entities, including, officially, central administration offices, responsible for nature and environment protection (TOKARSKA-GUZIK et al. 2015, TYKARSKI 2015). There is also a range of projects and initiatives gathering natural data, classified as citizen science[8], which encompass some individual activities, including also scientific employees of particular institutions. It is noteworthy that the largest projects in this group, such as www.inaturalist.org or www.ebird.org closely collaborate with GBIF, making their data globally available.

In 2010-2012, the project "Biodiversity Map of Poland in the global integration system on biodiversity data[9] was realized within the KSIB network; the project comprised intensive digitization works, which resulted in creation of a system integrating data from different kinds of sources: collections of specimens, bibliography, notes, photo images, etc. The scope of the data encompassed selected groups of insects (Coleoptera, Hemiptera, Lepidoptera), serving as a model for testing integration methods and presenting complex content (taxonomy, occurrence records, maps, bibliography, collections etc.), and the resulting tools were meant to be used at the national level. More than 100 people were involved in creation of the database, a significant part of this group consisted of KSIB members, which generated new ties in the local entomological community. The Biodiversity Map portal successfully implemented all the basic assumptions,

[6]www.gbif.org/dataset/search?publishing_country=PL
[7]GBIF portal in its present form does not show the date of the initial joining of a base; earlier activities may be viewed by Internet Archive Wayback Machine.
[8]For example, the Polish www.lepidoptera.eu, www.ornitho.ch (with local partners in many countries), www.observation.org and many others
[9]www.biomap.pl

in the following years expanding its capabilities and scope, including new data, also from outside the original three insect orders (over 1 million occurrence records in total). BioMap Connect, a module for entering and editing data from the browser level, has been operating since 2018 (not planned in the original scope of the project).

This experience gained in this way led to the creation of the concept of integrating the rich data resources of KSIB, presently implemented through the IMBIO project in the framework of the Digital Poland Operational Programme. The Programme, though formally not having scientific character, provides significant funds for digitization activities to more than 20 Polish scientific centers collecting natural data, which are presently carrying out several projects[10] of a similar nature simultaneously. It is worth paying attention to the fact that most of them opted for the GBIF as an external platform for sharing natural data (external repository), so that, irrespectively of the extent of integration of the systems on the country level, the output data of the digitization activities of all institutions involved will have a common access platform. Since its very beginning, GBIF has been acting according to FAIR rules (long before this term was coined), becoming a universal and increasingly used medium for biological data of the organismal level, connected with occurrence of species. Its choice, made independently by numerous institutions, is a step in the right direction. In this way, national data enter the global information flow, and the institutions are included in the group of units that make up the Network. Currently (June 2021), GBIF provides more than 1.8 billion data on the distribution of organisms, of which 180 million are for specimen collections. The mutual relations of the national data sources discussed here, in the context of presence in GBIF, are presented in Fig. 2.

The IMBIO project is the largest Polish undertaking of digitization of natural data,

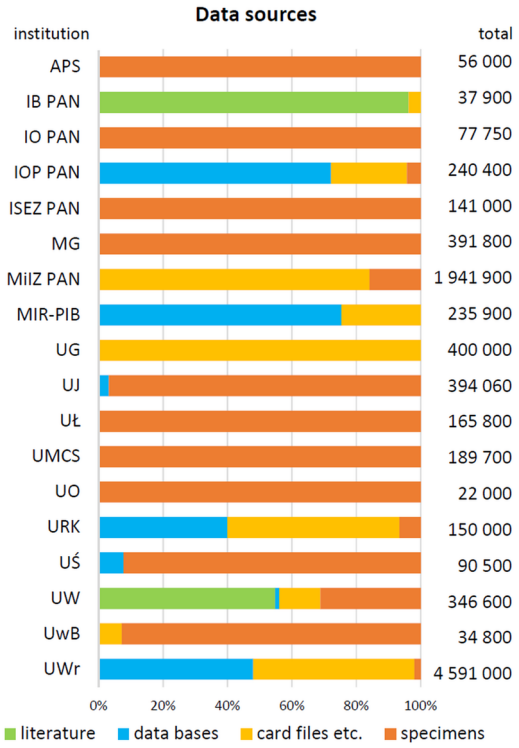

Fig. 3. Data planned to be digitalized within the IMBIO project, divided as to the type of records (specimens/observations), and the main groups of organisms. The presented values pertain to the number of recordings – distribution records, defined as a taxon-place-time relation.

in terms of scope, complexity and abundance of processed content. 18 institutions are its participants, with the University of Warsaw as a leader (Fig. 3). The project plans to deliver 9.5 million occurrence records, of which over 2 million will be based on specimens deposited in collections, and the remaining ones are connected with other types of scientific sources: index files, published and unpublished materials, publicly unavailable older databases (Fig. 3, 4). Including these resources in one project will allow their technical integration: all their content will be available literally at your fingertips. Their external access platform will be GBIF and related systems; on the national ground, a new version of the Biodiversity Map portal is being built with a wide range of filters and methods of searching complex content.

We are undertaking a huge challenge, due to its organizational complication, the level of data complexity and expectations as to its functionality, which means usefulness of the construed system. The success of this plan may highly improve the recognition and importance of units responsible

[10]Projects of the Operational Programme Digital Poland, connected with national natural data:
· AMU Nature Collections - online (AMUNATCOLL): digitization and sharing of natual data collection of the Faculty of Biology of Adam Mickiewicz University in Poznań (2018)
· OZwRCIN – Open Resources of the Digital Repository of Scientific Institutes (2018)
· e-Puszcza. Podlasie Region digital repository of natural scientific data (2019)
· Integrated virtual Herbarium of Pomerania Herbarium Pomeranicum – digitization and sharing of herbarium collections of university units of Pomerania by connection and digital sharing thereof (2019)
· Integration and mobilization of data on biotic diversity of Eukaryota in resources of Polish scientific institutions (2020).

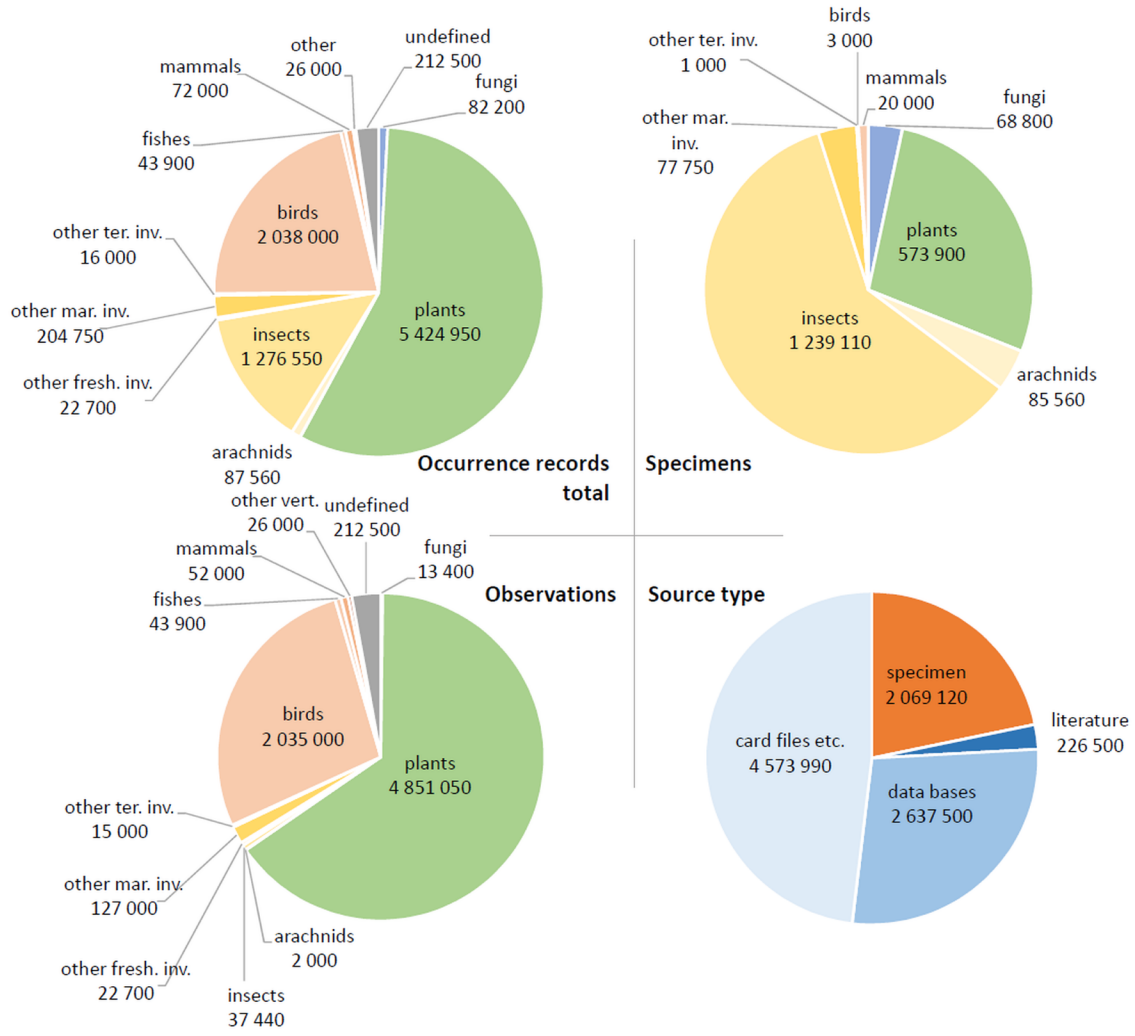

Fig. 4. Data planned to be digitalized within the IMBIO project, according to the partners of the project and the type of source (item collections, literature, databases, indexes, notes, etc.).

The presented values show the number of recordings – distribution records, defined as a taxon-place-time relation. Abbreviations: APS – Pomeranian Univ. in Słupsk, IB PAN – W. Szafer Inst. of Botany PAS, IO PAN – Inst. of Oceanology PAS, IOP PAN – Inst. of Nature Conservation PAS, ISEZ PAN – Inst. of Systematics and Evolution of Animals PAS, MG – Upper Silesian Museum, MIR-PIB – National Marine Fisheries Research Inst., MiIZ PAN – Museum and Inst. of Zoology PAS, UG – Univ. of Gdańsk, UJ – Jagiellonian Univ., UŁ – Univ. of Lodz, UO – Univ. of Opole, UMCS – Maria Curie-Skłodowska Univ., URK – Univ. of Agriculture in Krakow, UŚ – Univ. of Silesia in Katowice, UW – Univ. of Warsaw, UwB – Univ. of Bialystok, UWr – Wrocław Univ.

for natural history research. I truly believe that the articles by the participants of the project presented in this issue of KOSMOS will serve as a perfect illustration of the depth and complexity of this subject, and will allow readers to realize how valuable and remarkable are natural data resources, gathered and preserved by amazing people inhabiting the land "from the Baltic Sea to the Tatra Mountains".

REFERENCES

BAKKER F.T., ANTONELLI A., CLARKE J.A., COOK J.A., EDWARDS S.V., ERICSON P.G.P., FAURBY S., FERRAND N., GELANG M., GILLESPIE R.G., IRESTEDT M., LUNDIN K., LARSSON E., MATOS-MARAVÍ P., MÜLLER J., VON PROSCHWITZ T., RODERICK G. K., SCHLIEP A., WAHLBERG N., WIEDENHOEFT J., KÄLLERSJÖ M., 2020. *The Global Museum: natural history collections and the future of evolutionary science and public education*. PeerJ. 8, doi: 10.7717/peerj.8225.

DASZKIEWICZ P., IWAN D., KOWALSKI H., MIERZWA-SZYMKOWIAK D., ZABOROWSKI R. 2016. *Fedorowicz Z., Felisiak S. 150-lecie Gabinetu Zoologicznego w Warszawie (1818-1968)*. Memorabilia Zool. New Ser. 1, 35-38.

FACHIN D. A., COURI M. S., DE MELLO-PATIU C. A., 2016. *An illustrated catalogue of the types of Stratiomyidae (Diptera: Brachycera) in the collection of Museu Nacional, Rio de Janeiro, Brazil*. Zootaxa 4084, 361-376.

HEDRICK B., HEBERLING M., MEINEKE E., TURNER K., GRASSA C., PARK D., KENNEDY J., CLARKE J., COOK J., BLACKBURN D., EDWARDS S., DAVIS C., 2020. *Digitization and the Future of Natural History Collections*. BioScience, doi: 10.1093/biosci/biz163.

GRESHKO M., 2018. *Fire Devastates Brazil's Oldest Science Museum. National Geographic*. https://www.nationalgeographic.com/science/article/news-museu-nacional-fire-rio-de-janeiro-natural-history.

LANNOM L., KOUREAS D., HARDISTY A., 2019. *FAIR Data and Services in Biodiversity Science and Geoscience*. Data Intelligence 2, 122-130.

MEGA E. R., 2020. *Second Brazilian museum fire in two years reignites calls for reform*. Nature 583, 175-176.

MILLER S., BARROW L., EHLMAN S., GOODHEART J., GREIMAN S., LUTZ H., MISIEWICZ T., SMITH S., TAN M., THAWLEY C., COOK J., LIGHT J., 2020. *Building Natural History Collections for the Twenty-First Century and Beyond*. BioScience, doi: 10.1093/biosci/biaa069.

MOON P., 2018. *A Brazilian mourns what was lost in the National Museum fire*. Mongabay, https://news.mongabay.com/2018/09/a-brazilian-mourns-what-was-lost-in-the-national-museum-fire.

SANTOS D., RIBEIRO G. C., CABRAL A. D., SPERANÇA M. A., 2018. *A non-destructive enzymatic method to extract DNA from arthropod specimens: Implications for morphological and molecular studies*. PLoS One 13, doi: 10.1371/journal.pone.0192200.

TOKARSKA-GUZIK B., CHYBIORZ R., PARUSEL J. B., 2015. *Baza Danych Przestrzennych w zarządzaniu zasobami środowiska przyrodniczego Województwa Śląskiego*. Uniwersytet Śląski w Katowicach, Katowice.

TYKARSKI P., 2015. *Ogólnodostępne systemy gromadzenia danych o różnorodności biologicznej i możliwości ich wykorzystania*. [W:] *Baza Danych Przestrzennych w zarządzaniu zasobami środowiska przyrodniczego Województwa Śląskiego*. TOKARSKA-GUZIK B., CHYBIORZ R., PARUSEL J. B. (red). Uniwersytet Śląski w Katowicach, Katowice. 53-65.

WANDELER P., HOECK P. E. A., KELLER L. F., 2007. *Back to the future: museum specimens in population genetics*. Trends Ecol. Evol. 22, 634-642.

WEBSTER M. S., 2017. *The Extended Specimen: Emerging Frontiers in Collections-based Ornithological Research: Emerging Frontiers in Collections-Based Ornithological Research*. CRC Press, doi: 10.1201/9781315120454.

WIPFLER B., POHL H., YAVORSKAYA M. I., BEUTEL R. G., 2016. *A review of methods for analysing insect structures — the role of morphology in the age of phylogenomics*. Curr. Opin. Insect Sci. 18, 60-68.

Piotr Tykarski

Fundusze Europejskie Polska Cyfrowa
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego