

Piotr TYKARSKI (Warszawa)

BAZY DANYCH I POZIOM ORGANIZMALNY BIORÓŻNORODNOŚCI

Dane przyrodnicze — trochę historii...

Wiedza o gatunkach żyjących na Ziemi gromadzona była od samego początku interesowania się ludzi tym, co ich otacza. Początkowo zapewne miała ona przede wszystkim ułatwiać przetrwanie — dotyczyła tego, które gatunki są jadalne, które niebezpieczne, a które lecznicze. Zdobywano ją wielokrotnie i niezależnie w różnych rejonach świata. Do czasu pojawienia się pisma mogła być przekazywana jedynie ustnie, z pokolenia na pokolenie, i wygasała wraz z zanikiem grupy, która ją posiadała. Nie sposób wskazać momentu, kiedy zaczęto zbierać i utrzymywać wiadomości przyrodnicze, które nie miały bezpośredniego zastosowania praktycznego. Tak się jednak działo, czego dowody mamy w pracach starożytnych filozofów. Oprócz gromadzenia i utrwalania wiedzy w formie pisanej, zaczęto także już wtedy tworzyć pierwsze zbiory okazów przyrodniczych przeznaczonych do studiowania natury. Jedną z pierwszych wielkich kolekcji był zapewne zbiór Arystotelesa z okresu trwania jego Liceum (założonego ok. 335 r. p.n.e.). Nawet jeśli jej istnienie nie jest dostatecznie mocno udokumentowane, to faktem są bardzo bogate i wnikliwe traktaty o bio-

logii roślin i zwierząt, które wśród wielu dzieł pozostawił po sobie ten wybitny filozof. Arystoteles jako jeden z pierwszych, a na pewno najbardziej znany ze starożytnych przyrodników, gromadził wiedzę o organizmach w sposób metodyczny i starał się je rozróżniać na podstawie wielu cech, grupowanych w rozmaite kategorie, co dało początek taksonomii. Dzięki kopiowaniu manuskryptów, tłumaczeniom na łacinę i współczesne języki oraz wielokrotnym wznowieniom w postaci drukowanej jego prace przetrwały do dziś.

Dokonania Filozofa są istotne przede wszystkim dlatego, że pozostawiona przez niego wiedza stanowi pewnego rodzaju system, uporządkowany układ informacji. Dzięki swojej klarowności system ten stał się przydatny dla wielu późniejszych badaczy, którzy osiągnięcia Arystotelesa mogli rozwijać, udoskonalać i uzupełniać nowymi odkryciami. Musiało minąć dwa tysiące lat, by powstało dzieło o równie istotnym wpływie na sposób myślenia następnych pokoleń. Mowa oczywiście o „Systema Naturae” szwedzkiego przyrodnika Linneusza, twórcy nowożytnej taksonomii. Drugim ważnym spostrzeżeniem, nasuwającym się przy analizie historii dzieł przyrodniczych Arystotelesa, jest sam fakt ich wielokrotnego kopiowania. Jest to znakomity

przykład tego, że cenne zbiory danych mogą przetrwać wieki, choć ich forma współczesna daleko odbiega od oryginalnej. Arystoteles nie znał przecież papieru i technologii druku, a pisał i mówił starożytną greką.

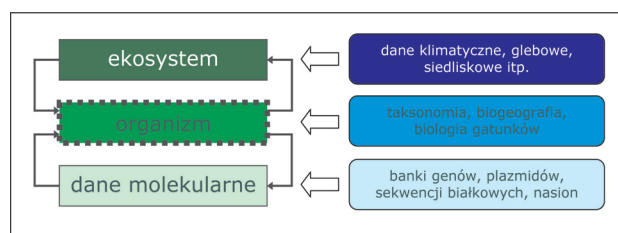
... i współczesność

Od czasów Linneusza nauki przyrodnicze nabrały charakteru wiedzy uporządkowanej, która wraz z upływem lat zaczęła przyrastać w coraz większym tempie. Współczesna taksonomia w dalszym ciągu wykorzystuje założenia szwedzkiego badacza i nadal pełni rolę szkieletu spinającego wszystkie dziedziny biologii. Pierwotne jednostki systematyczne (gromada, rząd, rodzina, rodzaj, gatunek, odmiana¹) musiały zostać uzupełnione o dodatkowe i pośrednie, pozwalające lepiej odwzorować skomplikowany, hierarchiczny układ, porządkujący bogactwo form życia. Rozwój nauk biologicznych wymusza ciągłe zmiany w podziale systematycznym organizmów, tak aby jak najlepiej oddawać pokrewieństwa pomiędzy nimi. I choć wiadomo, że jest to układ sztuczny i niedoskonały, wymagający nieustannych poprawek, to na całym świecie uznaje się go jako podstawę, której wygodniej używać niż alternatywnych propozycji ujęcia wyższych kategorii taksonomicznych, opartych np. na założeniach kladystyki².

Istnienie systemu taksonomicznego jest niezbędne przy wszelkich próbach odwzorowania zjawisk przyrodniczych. Dane uzyskane w wynikach jakichkolwiek badań odnoszą się przecież do konkretnych organizmów, a te zawsze należą do jakiegoś gatunku, nawet jeśli nie został on jeszcze opisany i nie posiada nazwy. Wyższe jednostki taksonomiczne są względne i podlegają zmianom, będąc jedynie obrazem hierarchii pokrewieństw. Gatunki natomiast istnieją realnie, jako populacje osobników mogących krzyżować się między sobą, izolowane rozrodczo od innych populacji i zajmujące określoną niszę ekologiczną³. Niezależnie od tego, czy mówimy o przepływie energii przez ekosystem czy o translacji genów mitochondrialnych, to opisywane zjawiska wiążą się z określonymi gatunkami.

Dochodzimy tu do definicji pojęcia bioróżnorodności (p. artykuł „Kłopoty z bioróżnorodnością” w tym zeszycie). Jest to termin stosunkowo nowy, powstały pod koniec XX w., gdy zdano sobie sprawę z szybkiego tempa utraty zasobów przyrodniczych, skutkującego zniknięciem wielu gatunków i nieodwracalnym spadkiem różnorodności form wykształconych przez miliony lat ewolucji. Pojęcie to ma szerokie znaczenie, odnosząc się zarówno do zróżnicowania biocenozy i ekosystemów, jak i bogactwa alleli i

zasobów genetycznych w populacji. Ogniwem łączącym te skrajne aspekty jest osobnik i gatunek. Można więc mówić o trzech składowych, czy też poziomach bioróżnorodności: wewnątrzgatunkowym — molekularnym, gatunkowym — organizmalnym oraz ponadgatunkowym — ekosystemowym (ryc. 1). Do bioróżnorodności rozumianej w ten sposób znajdujemy odniesienia niemal we wszystkich działach biologii, gdyż prawie każdy rodzaj dyscypliny biologicznej wiąże się z którymś z jej poziomów. Najczęściej jednak pojęcie to zawęża się do poziomu organizmalnego.



Ryc. 1. Poziomy bioróżnorodności i powiązane z nimi kategorie danych

Ilość informacji gromadzonych każdego dnia przez naukowców zajmujących się naukami przyrodniczymi jest tak duża, że dawno już przestała się mieścić na kartkach papieru; często zresztą, a może nawet już z reguły, do zapisu bezpośrednich wyników badań konieczne jest zastosowanie innych nośników informacji. Najpopularniejszymi dziś środkami do przechowywania i przetwarzania danych są oczywiście komputery i programy komputerowe. Z uwagi na ilość i stopień komplikacji informacji, które gromadzimy, potrzebne są odpowiednie narzędzia do sprawnego zarządzania i wykorzystania danych. Elektroniczne bazy danych zaczęto wykorzystywać w naukach przyrodniczych już w latach sześćdziesiątych XX w., ale obecnie ich znaczenie i zakres zastosowań znacznie wzrosły. Upowszechnienie się Internetu otworzyło drogę do nowej dziedziny — wykorzystania sieci komputerowych do wymiany i łączenia informacji o bioróżnorodności pochodzących z różnych źródeł. To, co kiedyś było w sferze marzeń przyrodników — łatwy i szybki dostęp do specjalistycznej wiedzy — dziś staje się faktem. Postępy informatyki w zakresie metod umożliwiających opracowywanie i analizę danych z zakresu poziomu organizmalnego bioróżnorodności są tak znaczne, że zaczęto używać określenia, który nie ma jeszcze polskiego odpowiednika — „biodiversity informatics”. Istniejący już termin „bioinformatyka” stosowany jest z reguły w odniesieniu tylko do danych z poziomu molekularnego. Uzasadnione więc wydaje się rozszerzenie go rów-

¹ Pierwotne, linneuszowskie znaczenie tej kategorii zostało zastąpione bardziej ścisłym pojęciem podgatunku. Współcześnie odmiana jest jednostką umowną, stosowaną zwykle na określenie barwnych lub hodowlanych form gatunku.

² Kladystyka — kierunek w taksonomii, w którym podstawowym założeniem jest, że w trakcie ewolucji gatunków są w stosunku do siebie grupami siostrzanymi o tej samej randze. Z tego punktu widzenia większość wyższych jednostek taksonomicznych okazuje się nierównocenna. Np. ptaki są grupą siostrzaną dla części dinozaurów, bo dane paleontologiczne wskazują, że wywodzą się z tej grupy. Jeśli więc potraktować gady jako zbiór gatunków pochodzących od jednego przodka, obejmujący m.in. żółwie, łuskonośne i hatterie, to należałoby do nich włączyć także ptaki. Do tego podejścia nie przystaje tradycyjny podział kręgowców na pięć gromad.

³ Nie ma prostej i doskonałej definicji gatunku. Przytoczona powyżej to współczesna, sformułowana przez Ernsta Mayra w 1982 r. Najwięcej problemów sprawiają organizmy jednokomórkowe, które mogą rozmnażać się bezpłciowo, oraz rośliny, których gatunki o wiele częściej się krzyżują niż zwierzęta, dając płodne potomstwo. Nie zmienia to faktu, że w ogromnej większości przypadków koncepcja gatunku sprawdza się bardzo dobrze.

niez na wyższe szczeble organizacji świata żywego (ramka 1).

Ramka 1 — Definicja bioinformatyki

„Badania, rozwijanie oraz zastosowania narzędzi i metod obliczeniowych w celu szerszego wykorzystania danych biologicznych, medycznych, behawioralnych i dotyczących zdrowia, włączając w to metody uzyskiwania, przechowywania, organizowania, archiwizowania, analizy i wizualizacji takich danych.”

oryginał: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

źródło: Working Definition of Bioinformatics and Computational Biology — July 17, 2000, National Institute of Health (USA), www.bisti.nih.gov

Trochę szczegółów technicznych

Czym jest „baza danych”? Określenie to jest często nadużywane i nie do końca właściwie interpretowane. W potocznym znaczeniu taką nazwę może otrzymać każdy rodzaj zbioru informacji, zwłaszcza gdy ma układ katalogu lub tabeli. Zdarza się więc, że miano bazy danych uzyskują np. wykazy gatunków zapisane w postaci kolejnych linii tekstu (ryc. 2A). O wiele bliższy bazie danych w ujęciu informatycznym jest zbiór informacji przechowywanych w formie arkusza kalkulacyjnego (np. w popularnym programie MS Excel), dającego dużą wygodę w organizowaniu i analizie danych (ryc. 2B). Możliwości programów tego typu znacznie ułatwiają ujednolicenie formatu zapisu, kopowanie i operacje na całych blokach informacji.

W pełni funkcjonalna baza danych to jednak coś więcej. Choć nie ma jednej powszechnie przyjętej definicji bazy danych, a rodzajów aplikacji jest wiele, to zdecydowaną większość z nich łączy fakt istnienia określonej struktury danych. Typem bazy, który najbardziej się upowszechnił i może stanowić wspaniałe narzędzie wspierające pracę w dziedzinach takich jak nauki przyrodnicze, są bazy relacyjne. Prostsze i łatwiejsze do opanowania dla osób bez przygotowania informatycznego są np. MS Access, OpenOffice.org Base czy Kexi (te dwie ostatnie to produkty darmowe, typu *open source*). Do bardziej zaawansowanych, dużo potężniejszych systemów należą między innymi PostgreSQL, MySQL, Oracle.

Oto zasadnicze cechy takiej bazy:

Dane zorganizowane są w formie tabel, wzajemnie ze sobą powiązanych. Powiązania te nazywamy relacjami (ryc. 2C).

Podstawową jednostką informacji, z których składa się tabela, jest rekord, czyli wiersz, w którym dane porozidzielane są na poszczególne pola, widoczne jako kolumny.

Do obliczeń, analiz i wyświetlania danych w różnych układach służą tzw. zapytania (inaczej kwerendy, od ang. *query*, lub widoki, od *view*), umożliwiające operacje na wszystkich jednocześnie lub wybranych według zadanych kryteriów rekordach tabeli (ryc. 2D, 3).

Najistotniejsze dla sprawnego posługiwania się tym narzędziem jest przyswojenie sobie odmiennej „filozofii”

działania bazy relacyjnej w stosunku do tabel takich jak arkusz kalkulacyjny. Zasadnicza różnica polega na tym, że tu jednostką informacji jest nie wartość w pojedynczej komórce, ale od razu cały wiersz wraz ze wszystkimi polami. Usunięcie informacji oznacza często usunięcie całego rekordu. Jeśli dane w tabeli sortujemy według wartości w jakimś polu (kolumnie), aby ułożyć je np. w porządku alfabetycznym, sortowane są od razu całe rekordy, a nie tylko wybrane pole.

Zapytania do bazy dają możliwość oglądania tych samych danych w różnych układach, tworzenia błyskawicznych podsumowań itp., co w standardowym arkuszu kalkulacyjnym wymaga znacznie więcej czasu i pracy. Podczas gdy w arkuszu układ tabeli jest powiązany z jej treścią i zmiana kolejności wierszy czy kolumn może łatwo wpłynąć na obliczenia, to w bazie relacyjnej realizowane jest rozdzielenie warstwy danych od warstwy obliczeń. Taką bazę można wyobrazić sobie jako wielowymiarową „chmurę informacyjną”, na którą możemy patrzeć z różnych stron poprzez stosowanie widoków (ryc. 3). Widok „spłaszcza” bazę do ustalonej w zapytaniu liczby wymiarów, pozwalając przy okazji na obliczenia. Wynik zapytania odpowiada temu, co dostępne jest w klasycznym arkuszu kalkulacyjnym, jednak nie wpływa na strukturę danych w bazie.

Poprawnie zbudowana baza (ramka 2) daje użytkownikowi duże możliwości. Przechowywanie danych różnych kategorii w różnych tabelach sprawia, że o wiele łatwiej jest zachować porządek i uniknąć błędów. Mechanizm relacji i możliwość „dowiązywania” nowych tabel oznacza, że łatwo można kojarzyć ze sobą różne informacje i rozszerzać bazę o nowe zasoby. Z kolei sprawne narzędzia wybierania i manipulacji rekordami, jakimi są kwerendy, pozwalają na błyskawiczne wyszukiwanie danych w oparciu o zadane kryteria, czyli wartości pól w powiązanych rekordach.

Ramka 2 — Cechy dobrze skonstruowanej bazy danych

— Rekordy w tabeli nie powinny się powtarzać; to oznacza, że przynajmniej jedno pole powinno zawierać wartości unikalne, występujące tylko w jednym rekordzie. Takie pole nosi nazwę klucza pierwotnego (ang. *primary key*) – np. identyfikator (zwykle unikalny numer) gatunku w liście nazw gatunkowych

— Poszczególne tabele powiązane są ze sobą poprzez odnośniki, zwykle w formie wartości liczbowych, wskazujące na związane ze sobą rekordy. Odnośnikiem do rekordów w innej tabeli są wartości klucza pierwotnego, które w tabeli powiązanej noszą nazwę klucza obcego (ang. *foreign key*) – np. odnośnik do tabeli z nazwami gatunków w tabeli obserwacji terenowych. Takie powiązania noszą nazwę relacji

— Każda tabela przechowuje inną kategorię danych (np. lista gatunków, typy habitatów, wyniki pomiarów badanych osobników danego gatunku)

— W tabeli powinno być jak najmniej pustych pól. Jeśli jest ich dużo, to może oznaczać, że dane pole powinno być wydzielone do osobnej tabeli

Tworzenie baz danych i związana z nim metodologia wykracza poza obszar wiedzy czysto technicznej, bo doty-

data	miejsce	nr porz	gatunek	4	samiec	samica	młode
2005-06-21	Rogów	1	zięba		1	1	1
		2	bogotka		1		
		3	pleszka		1		
		4	kos		1		
		5	pelczacz leśny				2
		6	zięba		1		
		7	bogotka				2
		8	kos		1		
2005-06-23	Ignaców	1	kos		1		
		2	szpak		1		
		3	szpak				5
		4	zięba		1		
		5	bogotka		1		
		6	kowalik				2
		7	pelczacz leśny				1
		8	szpak				5
		9	zięba		1		
2005-06-25	Brzeźnica	1	bogotka				4
		2	rudzik		1		
		3	kowalik				1
		4	bogotka		1		
		5	dzięciołek				
		6	rudzik				1

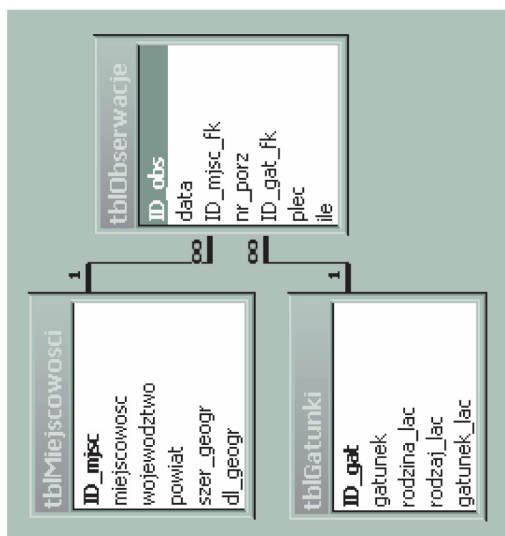
C

mięjscość	województwo	powiat	wsp. geogr.
Rogów	łódzkie	brzeziński	N 51,83; E 19,89
Ignaców	mazowieckie	grójce	N 51,77; E 20,64
Brzeźnica	świętokrzyskie	konecki	N 51,27; E 20,48

B

Czas	Miejsce	nr	gatunek	l. osobników	pleć
...					
2005-06-21	Rogów, woj. łódzkie, pow. brzeziński				
N 51,83; E 19,89					
1.	zięba: 1 m (samiec), 1 f (samica), 1 j (młode)				
2.	bogotka: 1 m				
3.	pleszka: 1 m				
4.	kos: 1 m, 1 f				
5.	pelczacz leśny: 2 j				
6.	zięba: 1 m				
7.	bogotka: 2 j				
8.	kos: 1 f, 1 m				
2005-06-23	Ignaców, woj. mazowieckie, pow. grójce				
N 51,77; E 20,64					
1.	kos: 1 m				
2.	szpak: 1 m, 1 f				
3.	szpak: 5 j				
4.	zięba: 1 m				
5.	bogotka: 1 m				
6.	kowalik: 2 j				
7.	pelczacz leśny: 1 j				
8.	szpak: 5 j				
9.	zięba: 1 f, 1 m				
2005-06-25	Brzeźnica, woj. świętokrzyskie, pow. konecki				
N 51,27; E 20,48					
1.	bogotka: 4 j				
2.	rudzik: 1 m				
3.	kowalik: 1 j				
4.	bogotka: 1 m				
5.	dzięciołek: 1 j				
6.	rudzik: 1 m				

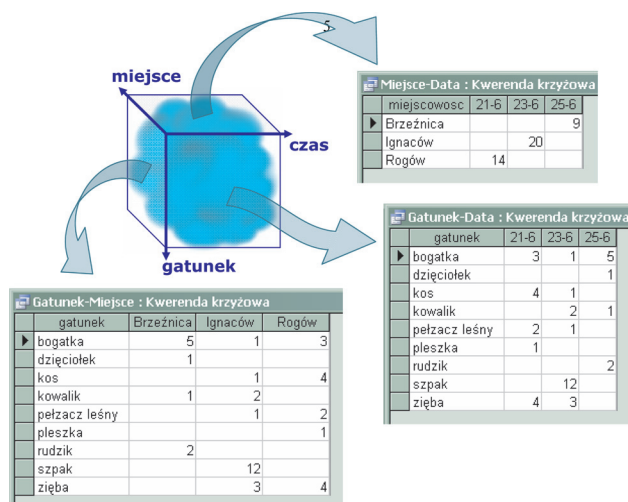
A



ID_obs	data	miejscowosc	województwo	powiat	nr_porz	rodzina_lac	rodzaj_lac	gatunek_lac	gatunek	szer_geogr	dl_geogr	ile	plec
1	2005-06-21	Rogów	łódzkie	brzeziński		1 Fringillidae	Fringilla	coelebs	zięba	51.83	19.89	1	m
2	2005-06-21	Rogów	łódzkie	brzeziński		1 Fringillidae	Fringilla	coelebs	zięba	51.83	19.89	1	f
3	2005-06-21	Rogów	łódzkie	brzeziński		1 Fringillidae	Fringilla	coelebs	zięba	51.83	19.89	1	j
13	2005-06-23	Ignaców	mazowieckie	grójce		1 Turdidae	Turdus	merula	kos	51.77	20.64	1	m
24	2005-06-25	Brzeźnica	świętokrzyskie	konecki		1 Paridae	Parus	major	bogotka	51.27	20.48	4	j
4	2005-06-21	Rogów	łódzkie	brzeziński		2 Paridae	Parus	major	bogotka	51.83	19.89	1	m
14	2005-06-23	Ignaców	mazowieckie	grójce		2 Sturnidae	Sturnus	vulgaris	szpak	51.77	20.64	1	m
15	2005-06-23	Ignaców	mazowieckie	grójce		2 Sturnidae	Sturnus	vulgaris	szpak	51.77	20.64	1	f
25	2005-06-25	Brzeźnica	świętokrzyskie	konecki		2 Saxicolidae	Erithacus	rubicula	rudzik	51.27	20.48	1	m
5	2005-06-21	Rogów	łódzkie	brzeziński		3 Saxicolidae	Phoenicurus	phoenicurus	pleszka	51.83	19.89	1	m
16	2005-06-23	Ignaców	mazowieckie	grójce		3 Sturnidae	Sturnus	vulgaris	szpak	51.77	20.64	5	j

Ryc. 2. Przykładowy zbiór obserwacji ptaków leśnych zapisany w różnych formatach. Warto zwrócić uwagę, że zachowana zostaje oryginalna kolejność (numer) obserwacji; są to informacje dodatkowe o charakterze metadanych (informacja o informacji), jednak nie należy ich tracić. A — tekst, w układzie zbliżonym do formy zapisu w terenie, wg ustalonego szablonu. B — arkusz kalkulacyjny — forma czytelniejsza i łatwiejsza w obróbce; zwraca uwagę konieczność utworzenia dodatkowej tabeli dla danych o miejscach obserwacji. C — jeden z możliwych układów tych samych danych w bazie relacyjnej; trzy tabele dla trzech typów danych, połączone relacjami. W tabeli nazw gatunkowych i miejscowości nie ma powtórek, a w tabeli obserwacji pustych pól. D — wirtualna tabela — wynik zapytania do bazy (kwerendy), zestawiającego dane ze wszystkich trzech tabel i podającego wszystkie dane wiążące się z poszczególnymi obserwacjami. Sortowanie rekordów w tabeli jest całkowicie dowolne — w tym wypadku wykonano je wg pól nr_porz (numer obserwacji w terenie) i ID_obs (numer obserwacji w bazie).

czy nie tylko sposobów wydajnego zarządzania informacją, ale także podsuwa rozwiązania mogące zwiększyć możliwości analizowania otaczającej nas przyrody. Bazy danych z jednej strony zmniejszają ilość pracy niezbędną do gromadzenia informacji i ułatwiają ich wykorzystanie, a z drugiej strony pozwalają rozszerzyć ich zakres, tworząc połączenia pomiędzy różnymi dziedzinami.



Ryc. 3. Baza z ryc. 2C jako wielowymiarowa „chmura informacji” oglądana za pomocą zapytań, spłaszczających dane do prostych tabel przedstawiających sumaryczną liczbę obserwowanych ptaków w różnych ujęciach

Jak wygląda typowa baza danych dotycząca bioróżnorodności poziomu organizmowego? Nie ma jednego ustalonego standardu i zapewne nigdy go nie będzie, bo każdy autor stosuje własne metody i gromadzi odmienne informacje, dostosowując układ treści do swoich potrzeb. Niemniej większość tego typu danych zawiera przynajmniej trzy wspólne komponenty: informację o gatunku, miejscu i czasie, w którym go zarejestrowano. Mówiąc w skrócie, odpowiada na pytania: „co? gdzie? kiedy?”. Niezależnie od przeznaczenia danej bazy oraz innych przechowywanych w niej informacji, te trzy elementy są z reguły obecne. Różny bywa natomiast stopień ich dokładności.

Ustalony schemat struktury informacji jest podstawą umożliwiającą łączenie danych z różnych źródeł w jeden, większy system. Od momentu powstania Internetu idea ta była wielokrotnie wykorzystywana do tworzenia standardów wymiany informacji o bioróżnorodności i podejmowania mniej lub bardziej udanych prób budowania opartych o nie sieci komputerowych. Wśród istniejących obecnie, siecią o największym zasięgu, mającą duże zasługi w upowszechnianiu się tych rozwiązań, jest GBIF (*Global Biodiversity Information Facility* — www.gbif.org). Łączy ona dane pochodzące z wielu instytucji naukowych i organizacji na całym świecie. Rekordy informacji dostępne poprzez Portal GBIF pochodzą z wielu różnego typu baz, rozproszonych w różnych miejscach i mających różną budowę i funkcję (więcej o podstawach działania sieci wymiany informacji w artykule „Bioróżnorodność Ziemi — wspólna sprawa, wspólne działanie...”). Ponieważ jednak wszystkie używają tego samego układu danych „na wyjściu”, możliwe jest łączenie informacji pochodzących od wielu dostawców (ang. *data providers*).

Warto zwrócić uwagę, że „z punktu widzenia” systemu komputerowego nie ma znaczenia, czy rekord pochodzi z muzealnej kolekcji owadów, arkusza zielnikowego w herbarium czy z bazy obserwacji ptaków. Wszystkie trzy podstawowe elementy są obecne — można go więc umiejscowić w czasie i przestrzeni, a także wiadomo, jakiego organizmu dotyczy. Różna jest oczywiście waga takich danych. Rekordy związane z kolekcjami mają to do siebie, że można zwykle sprawdzić poprawność oznaczenia; okaz jest w takim przypadku dowodem rzeczowym. Tego elementu na ogół brak danym o charakterze obserwacyjnym, o ile nie istnieje dokumentacja dodatkowa w postaci zdjęcia, notatek, czy nagrania dźwiękowego.

Celem GBIF jest doprowadzenie do tego, aby jak najwięcej zasobów informacji o bioróżnorodności, rozproszonych na świecie, było dostępnych dla użytkowników Internetu w jednym systemie, będącym w stanie wysyłać zapytania do rozproszonych baz i przedstawiającym odpowiedzi w syntetycznej formie. Jest to niewątpliwie idea szczytna i mogąca przynieść wiele pożytku nie tylko nauce. Wiąże się z nią problemy natury politycznej, sprawa przestrzegania praw autorskich dostawców danych i sporo innych, czysto technicznych. Większość z nich już doczekała się rozwiązania lub nastąpi to wkrótce.

Jednak długo jeszcze będzie aktualny problem niedoboru samych danych, istoty całego systemu. Choć nowe dane

Ramka 3 — Polecane strony internetowe:

— www.gbif.org — Global Biodiversity Information Facility — sieć o zasięgu światowym, zapewniająca dostęp do danych o rozmieszczeniu i taksonomii gatunków z całego świata

— www.ksib.pl — Krajowa Sieć Informacji o Bioróżnorodności — polska organizacja współpracująca z GBIF, udostępniająca dane pochodzące z krajowych ośrodków

— www.faunaeur.org — Fauna Europaea — baza taksonomiczna europejskich gatunków zwierząt lądowych i słodkowodnych

— www.marbef.org/data/erms — European Register of Marine Species — baza taksonomiczna europejskich gatunków zwierząt morskich

— www.emplantbase.org — Euro+Med PlantBase — baza taksonomiczna europejskich gatunków roślin

— www.speciesfungorum.org/Names/Names.asp — Index Fungorum — baza taksonomiczna grzybów całego świata

— www.species2000.org — Species 2000 — baza taksonomiczna gromadząca dane z wielu źródeł dotyczących wszystkich grup organizmów — jeden z fundamentów Elektronicznego Katalogu Życia (*Electronic Catalogue of Life*)

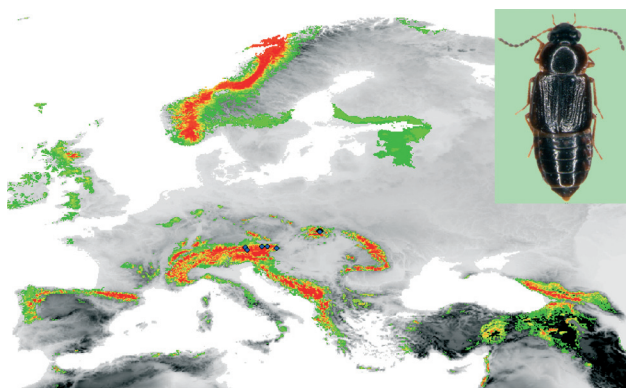
— www.fishbase.org — FishBase — dobry przykład bazy dotyczącej wszelkich aspektów biologii i biogeografii wybranej dużej grupy organizmów — ryb

— www.tolweb.org — Tree of Life — strona projektu poświęconego taksonomii wszystkich grup organizmów, zmierzającego do zebrania i aktualizacji wszelkich możliwych informacji o pokrewieństwach form życia na Ziemi

gromadzone są już zwykle w formie cyfrowej, niemal gotowej do włączenia w obieg, to wciąż pozostaje do wykonania mnóstwo pracy, aby dostosować te już istniejące, przechowywane w notatnikach, katalogach i publikacjach. Bardzo przydatna jest w tym celu znajomość zasad działania baz danych. Ich tworzenie nie wymaga wcale specjalistycznej wiedzy, za to czas zainwestowany w ich przygotowanie zwraca się wielokrotnie. Dotyczy to również sytuacji, gdy baza jest przeznaczona tylko do użytku wewnętrznego i ma służyć wyłącznie właścicielowi (co w przypadku danych naukowych zdarza się dość rzadko).

Zastosowania i perspektywy

O opłacalności inwestowania w tworzenie baz danych można się przekonać biorąc pod uwagę ich szczególną użyteczność w badaniach przyrodniczych i zarysowujące się perspektywy. Badania związane z ekologią, ochroną środowiska i spokrewnionymi naukami z reguły wymagają gromadzenia i łączenia wielu faktów, wyciągania wniosków na podstawie nakładania na siebie informacji z różnych źródeł. To jest właśnie pole, w którym sprawdzają się bazy danych.



Ryc. 4. Przykład modelowania niszy ekologicznej — ekstrapolacja rejonów występowania gatunku górskiego chrząszcza *Phyllodrepa crenata* na podstawie niepełnych danych o jego rozmieszczeniu (niebieskie punkty) i wybranych parametrów środowiska obszaru całej Europy (m.in. wysokość n.p.m., średnia temperatura roczna, ilość opadów). Obszary o większym prawdopodobieństwie wystąpienia gatunku zaznaczone na czerwono. Analiza wykonana przy pomocy programu GARP

Weźmy jako przykład analizę sieci troficznych i związków między elementami ekosystemu. Gatunki występujące w przyrodzie powiązane są wzajemnymi zależnościami, np. drapieżnik — ofiara, roślina — roślinożerca, pasożyt — żywiciel. Dość łatwo wyobrazić sobie bazę opisującą takie powiązania — np. tabelę „Roślinożercy”, której rekordy zawierają nazwy roślinożerców, tabelę „Rośliny Żywicielskie” oraz tabelę łączącą dwie poprzednie („Roślinożerca-Roślina”), zawierającą pary odnośników do obu powiązanych tabel. Mogą do tego dojść dodatkowe informacje, np. o częściach rośliny (liście, kwiaty, łodygi itp.), które są przez dany gatunek zjadane. Jeden rekord z tabeli roślinożerców może być związany z wieloma rekordami z

tabeli roślin i odwrotnie, z jednym gatunkiem rośliny może się wiązać wielu jej amatorów. W dużych zbiorach informacji orientacja w takiej masie wzajemnych powiązań może sprawiać niemałe kłopoty. Z poprawnie skonstruowanej bazy informacje takie wyciąga się natychmiast.

Im bardziej rozbudowana jest tego typu baza, tym bardziej zaczyna ona przypominać swego rodzaju model rzeczywistości. Porównanie to nabiera mocy, gdy informacje przyrodnicze zaczniemy wiązać z bazami GIS¹ (od *Geographic Information System*), odwzorowującymi środowisko geograficzne. Bazy tego typu służą do gromadzenia danych reprezentujących cechy środowiska fizycznego, pozwalając na analizy uwzględniające położenie obiektów względem siebie. Dane takie gromadzone są przez specjalistów różnych działów nauki — kartografów, geologów, czy klimatologów i z roku na rok stają się coraz powszechniej dostępne oraz dokładniejsze.

Jeśli baza o interesującej nas grupie gatunków zawiera odniesienia do położenia geograficznego, najlepiej w postaci współrzędnych geograficznych, to przy użyciu odpowiedniego oprogramowania można połączyć oba typy danych. Załóżmy, że posiadamy bogatą bazę informacji dotyczących występowania jakiegoś gatunku rośliny na danym obszarze i chcemy wiedzieć jak najwięcej o jego wymaganiach względem warunków środowiska. Jeśli nałożymy te dane na mapy glebowe badanego rejonu, mamy szansę zbliżyć się do poszukiwanej odpowiedzi. Jeśli są to badania większego obszaru, można kojarzyć dane o rozmieszczeniu gatunku z mapami klimatycznymi, wskazując preferowany przezeń zespół warunków klimatycznych (np. zakres temperatur, ilość opadów) (ryc. 4). Zamiast map rozkładu czynników środowiskowych można zastosować mapy rozmieszczenia gatunków powiązanych z interesującym nas organizmem, aby wytypować np. miejsca, gdzie potencjalnie mogłyby występować razem.

Metody obliczeniowe, pozwalające na powiązanie, a także wzajemne analizy obiektów biologicznych i geograficznych, to stosunkowo młoda dziedzina, która szybko się rozwija między innymi dzięki sieciom komputerowym, dostarczającym dużej ilości niezbędnych danych. Coraz więcej jest także odpowiednich aplikacji komputerowych. Dzięki temu możliwe jest modelowanie wymagań środowiskowych organizmu (tzw. modelowanie niszy ekologicznej, ang. *ecological niche modeling*) w oparciu o dostępne dane o jego występowaniu (dane punktowe) i cechy środowiska (określane przez właściwości elementów przestrzeni w systemie GIS). W ten sposób można na przykład przewidywać zmiany w rozmieszczeniu gatunku przy różnych scenariuszach zmian klimatycznych lub wyznaczać obszary, na których może on występować.

Powstaje więc właściwie nowa gałąź wiedzy, korzystająca z nowoczesnych narzędzi informatycznych, wykorzystujących zaawansowane algorytmy i rozwiązania programistyczne. Użycie terminu „bioinformatyka” jest tu jak najbardziej uzasadnione, tyle że w odniesieniu do innego poziomu bioróżnorodności. Bioinformatyka stosowana przez biologów molekularnych pozwala np. przewidzieć funkcję białka w komórce na podstawie znajomości jego sekwencji. Bioinformatyka z wyższego (w sensie organiza-

¹ Można spotkać również rzadziej spotykany skrót — SIP (System Informacji Przestrzennej)

cji świata żywego) poziomowi umożliwia poszukiwanie odpowiedzi na pytania o roli gatunku w środowisku, przewidywanie gatunków powiązanych czy zmian liczebności pod wpływem zmian w otoczeniu na podstawie znajomości jego obszaru występowania i preferencji ekologicznych. Zasadnicza różnica z punktu widzenia technologii informatycznych wykorzystywanych do tego celu polega na tym, że gromadzenie danych o występowaniu gatunków zachodzi o wiele wolniej niż np. sekwencjonowanie ich genomów. Stąd z reguły mamy do czynienia z niedoborem danych, na których opierają się analizy, i często modelowanie powiązań gatunku z parametrami środowiska zawiera elementy ekstrapolacji i rachunku prawdopodobieństwa. Weryfikacja poprawności wyników może z kolei wymagać badań terenowych, co wprowadza dodatkowe utrudnienia. Idea łączenia baz danych poprzez komputerowe sieci informacyjne tworzy szansę na pokonanie tej bariery. Jeśli tylko niezbędne dane na jakiś temat istnieją w sieci, można je pobrać w jednolitej formie i wykorzystać do analizy, choć pochodzą z różnych baz. W ten sposób połączenie zbiorów informacji cząstkowych jest czymś więcej niż tylko ich sumą i może otwierać drzwi do nowej wiedzy.

Zastosowanie technologii informatycznej do zbierania, przesyłania i przetwarzania danych przyrodniczych niesie ze sobą wielki potencjał do wykorzystania w działaniach praktycznych. Poza podstawową (co absolutnie nie znaczy, że mało istotną), najprostszą technicznie funkcją banków informacji o gatunkach, bazy danych o bioróżnorodności

mają szansę stać się elementami systemów monitoringowych. Wymaga to „jedynie” ich zdynamizowania i systematycznego zbierania danych oraz powiązania ich z gromadzonymi równolegle informacjami o środowisku. W ten sposób można na przykład śledzić i przewidywać występowanie gatunków inwazyjnych lub istotnych gospodarczo. Nałożenie na siebie informacji o rozmieszczeniu na danym obszarze gatunków chronionych, rzadkich lub zagrożonych pozwoliłoby uzyskać rodzaj mapy terenów cennych przyrodniczo i skonfrontować ją z istniejącą siecią obszarów chronionych w celu ich weryfikacji. Proste zestawienie punktów i powierzchni pomiarowych wykorzystanych w prowadzonych projektach naukowych może samo w sobie być pomocne w planowaniu przyszłych badań i ustaleniu „białych plam” — rejonów niezbadanych.

Przygniatająca większość tych wspaniałych możliwości rysujących się przed tą częścią bioinformatyki ma jeden zasadniczy mankament — do urzeczywistnienia potrzebuje paliwa, czyli danych. A do tego niezbędna jest świadomość potrzeby ich gromadzenia i znajomość metod, które to ułatwiają.

Wpłynęło 29.06.2007

Dr Piotr Tykarski jest adiunktem w Zakładzie Ekologii na Wydziale Biologii UW.

Zajmuje się ekologią owadów saproksylicznych a w KSIB odpowiada za organizację i udostępnianie baz danych.