Wystąpienia plenarne

Utilizing National and International Systems, Applications and Services for the Information Community

Robert D. Gelfeld

U.S. National Oceanographic Data Center/
World Data Center for Oceanography, Silver Spring
1315 East West Highway, Room 4230
Silver Spring, MD 20912
U.S.A
Tel: +1 301-713-3270
Robert.Gelfeld@noaa.gov

A key to utilizing national and international systems, applications, and services for the information community is the ability to provide general access to scientific data and information using current scientific innovations, research, and collaboration between science and society. The U.S. National Oceanographic Data Center (a component of the U.S. National Oceanic and Atmospheric Administration) is a national and international repository and dissemination facility for global oceanographic data. As such, the center acquires and preserves a historical record of the earth's changing environment to be used for operational applications and ocean climate research. Working cooperatively, the center provides data and information products and services to scientists, engineers, resource managers, policy makers, and other users around the world to document and describe the ocean's natural variability. It is an agency that enriches life through science and touches the lives of every person on the earth.

Integracja danych naukowych – modele i techniki

Tadeusz Pankowski

Politechnika Poznańska, Instytut Automatyki i Inżynierii Informatycznej, Pl. M.S.Curie 5, 60-965 Poznań
tadeusz.pankowski@put.poznan.pl

Streszczenie: Integracja danych naukowych staje się ważną dyscypliną badawczą łączącą metody i techniki baz danych, systemów zarządzania wiedzą, obliczeń rozproszonych oraz dziedzinowych dyscyplin naukowych – głównie nauk przyrodniczych, medycznych i technicznych. Specyfika danych naukowych i specyficzne wymagania dotyczące ich integracji, powodują rozwój nowych rozwiązań w tym zakresie. Tradycyjne metody integracji danych biznesowych rozszerzane są o możliwości wykorzystania ontologii dziedzinowych, technologii przepływów działań naukowych (ang. scientific workflows) i obliczeń w środowisku Grid.

Summary: Scientific data integration is becoming an important research field joining methods and techniques from all database technology, knowledge management, distributed computing and domain sciences – mainly natural science, medicine and technology. Specific features of scientific data and requirements considering its integration, have caused new solutions and developments in this field. Traditional approaches to integrating business data are extended towards using domain ontologies, scientific workflows and Grid computing.

1. Wstęp

Celem integracji danych jest umożliwienie użytkownikom (osobom i programom) sprawnego korzystania z danych zgromadzonych w różnorodnych heterogenicznych źródłach danych. Integracja danych jest obecnie jednym z głównych problemów badawczych i praktycznych w dziedzinie baz danych i systemów zarządzania wiedzą, choć dotąd dotyczyła głównie danych biznesowych. Jednak wraz ze wzrostem wydatków na badania w zakresie zarządzania danymi naukowymi, pojawia się nacisk na zwiększenie efektywności tych badań. Zaowocowało to rozwojem metod, narzędzi i systemów realizujących i wspierających procesy integracji danych naukowych [1][19]. Do podstawowych modeli integracji danych naukowych należą: (1) integracja materializująca (oparta na technologii hurtowni danych); (2) integracja wirtualna (zakładająca tworzenie schematu globalnego i odwzorowań między schematem globalnym i schematami lokalnych źródeł danych); (3) integracja z wykorzystaniem przepływów działań (ang. *workflows*) (oparta na konfigurowaniu procesów przepływu danych między źródłami danych i/lub programami obliczeniowymi). Pokazujemy, jak tradycyjne metody integracji danych biznesowych znajdują zastosowanie w integracji danych naukowych i jakie nowe problemy wymagają rozwiązań. W dyskusji odwołujemy się do różnych rozwiązań opracowywanych w ramach wielu projektów podejmowanych na świeci.

Do ważniejszych cech odróżniających integrację danych naukowych od integracji danych biznesowych można zaliczyć: (a) semantyczną i strukturalną różnorodność zarówno samych danych, jak i sposobów ich wykorzystywania; (b) bogatszy zestaw typów danych – oprócz dobrze sformatowanych danych relacyjnych i XML mamy do czynienia z dynamicznymi tablicami wielowymiarowymi o często ogromnych rozmiarach reprezentującymi obiekty graficzne i numeryczne, a także z obrazami oraz obiektami audio i video (które jednak ciągle pozostają poza głównym nurtem integracji semantycznej); (c) potrzeby sterowania przez użytkowników wieloma etapami procesu integracji związanymi z pozyskiwaniem danych, ich analizą, symulacją, weryfikacją i wizualizacją. Cechą wyróżniającą dane naukowe jest również występowanie różnego rodzaju *adnotacji* (ang. *annotations*) [2], które wprowadzane są przez ekspertów (a w tym przypadku właściwie każdy naukowiec jest ekspertem w swojej dziedzinie) i których integrowanie z danymi podstawowymi i odpowiednie udostępnianie w procesie integracji odgrywa bardzo ważną rolę. Z punktu widzenia oceny wiarygodności danych, ważną cechą danych naukowych jest ich *pochodzenie* (ang. *provenance*) [4], a informacje o pochodzeniu powinny być właściwie zarządzane, uzupełniane i przekazywane w procesie integracji.

2. Modele integracji danych

Integracja danych naukowych jest dużo trudniejsza niż integracja danych biznesowych [6][7][9][12]. W tym ostatnim przypadku integracji podlegają zasoby danych gromadzone w bazach danych (relacyjnych i XML-owych), a także dane semistrukturalne (takie jak poczta elektroniczna) i tekstowe. Wszystkie te rodzaje danych występują także w przypadku danych naukowych, ale dane naukowe charakteryzują się dużo większą różnorodnością, zarówno struktur danych, postaciami danych, a także ich znaczeniem (semantyką). W przypadku danych naukowych dużo bardziej złożone są też problemy dotyczące analizy, symulacji, weryfikacji hipotez i prognozowania [1].

2.1. Integracja materializująca

Najczęściej stosowaną metodą integracji danych biznesowych, a często także danych naukowych, jest ich gromadzenie w scentralizowanym repozytorium, a więc ich materializacja, czyli fizyczne zapamiętanie w jednej centralnej bazie danych zwanej wówczas *hurtownią danych* (ang. *data warehouse*) [6][9]. Podstawową techniką materializacji jest proces ETL (*Ekstrakcja/Transformacja /Ładowanie*) [13]. W procesie tym dane są pobierane z jednego lub z wielu źródeł danych, poddawane są transformacji zgodnie z opracowaną specyfikacją, a następnie są zapamiętywane w docelowej bazie danych. Jest to metoda używana od dawna również w kontekście danych naukowych. W ten sposób tworzonych jest wiele repozytoriów danych, które następnie udostępniane są w Internecie (przykładem jest system PIR, `http://pir.georgetown.edu`).

2.2. Integracja wirtualna

Integracja wirtualna polega na tworzeniu *schematu globalnego* (zwanego też *schematem mediującym*) i określeniu powiązań tego schematu ze schematami źródłowymi [23][16]. Cały system określany jest wówczas jako system *sfederowany* lub *federacja* systemów lokalnych. Podczas wykonywania poleceń użytkownik posługuje się schematem globalnym. Ma więc wrażenie, że istnieje jeden schemat obejmujący wszystkie możliwe dane istniejące w systemie sfederowanym, ma do dyspozycji pewien interfejs zapytań, który wspomaga go w procesie formułowania zapytań. W najprostszym przypadku zapytania są tworzone za pomocą odpowiedniej kombinacji słów kluczowych, a w najbardziej zaawansowanym – zapytania formułowane są względem ontologii opisującej cały system. Kolejnym zadaniem systemu jest określenie, jak rozłożyć zapytanie globalne na zapytania lokalne dotyczące poszczególnych lokalnych źródeł danych. Te operacje określane są jako *mediacje* – stąd też o schemacie globalnym mówi się także jako o *schemacie mediującym*. Lokalne źródła danych wyposażone są w *osłony* (ang. *wrappers*), których zadaniem jest przyjmowanie lokalnych zapytań, przekazywanie ich do wykonania w lokalnym systemie, a następnie zwracanie odpowiedzi zgodnie z modelem (formatem) zrozumiałym przez schemat mediujący [19]. W zależności od sposobu wykonywania zapytań formułowanych względem schematu globalnego (docelowego) mówimy o *wymianie danych* (ang. *data exchange*) i *reformułowaniu zapytań* (ang. *query reformulation*) [18].

Bardziej zaawansowaną metodą federacji są systemy *integracji P2P* (*Peer-to-Peer*). Wówczas nie ma jednego wyróżnionego schematu docelowego (mediującego), a rolę takiego schematu może pełnić każdy ze schematów związanych z poszczególnymi węzłami sieci (niezależnie czy są w nim zapamiętane również dane, czy nie). Wówczas zarówno wymiana danych, jak i reformułowanie zapytań propagowane są między węzłami zgodnie ze ścieżkami semantycznych powiązań określających odwzorowania między schematami [10][15][20]17].

Tego rodzaju integracja określana jest często jako integracja *federacyjna*, bo realizowana jest w systemie baz danych tworzących pewną *federację*. Federacja taka może być tworem dynamicznym, tzn. różne lokalne systemy mogą do niej wchodzić i z niej wychodzić [1].

Proponowane są różne podejścia mające ułatwić integrację wirtualną: przyjęcie standardowej terminologii jest jedną z głównych metod integracji treści poprzez ustalenie odwzorowań do standardowego słownictwa. Na przykład GeneOntology (`www.geneontology.org`) jest używana jako terminologia odniesienia przez 14 głównych zasobów danych, system TAMBIS ma własną ontologię TaO, podobnie DiscoveryLink czy SRS (`srs.ebi.ac.uk`). Szereg prac dotyczy integracji

w środowiskach P2P opartej na XML i usługach sieciowych Web Service, a także w środowiskach typu Grid Na przykład ComparaGRID (bioinf.ncl.ac.uk/comparagrid) realizuje integrację danych genetycznych dla badania zależności między genomami różnych gatunków. System implementowany jest jako warstwa pośrednia w środowisku Web/GRID, wspomagająca operacje nad bazami danych wyposażonymi w osłony z uwzględnieniem integracji danych poprzez odwoływanie się do słowników i ontologii odniesienia. Projekt Open Bio Foundation bioMOBY (www.biomoby.org) wspomaga interoperacyjność między centrami danych biologicznych i serwisami analitycznymi. Projekt NeuronBank (neuronbank.org) tworzy infrastrukturę informatyczną wspomagającą zarządzanie wiedzą o układach nerwowych. System jest federacją baz wiedzy, z których każda dotyczy jednego gatunku, opartą na zbiorze usług sieciowych i centralnym portalu udostępniającym operacje nad bazami danych.

2.3. Integracja za pomocą przepływów działań naukowych

W metodzie tej definiowanych jest szereg kroków realizujących specjalizowane działania cząstkowe. Poszczególne kroki organizowane są w sieć przepływu działań [11]. Utworzone, sprawdzone i zapamiętane przepływy działań mogą być wykorzystywane wielokrotnie przez różnych użytkowników pracujących w systemie integracji danych (przykładem takiego rozwiązania jest myExperiment, www.myexperiment.org).

Przepływy działań naukowych odgrywają coraz większą rolę w naukowej cyberprzestrzeni. Umożliwiają one "sklejanie" usług, zarządzanie danymi, analizy, symulacji i wizualizacji danych – operują często na ogromnych oraz złożonymi (strukturalnie i semantycznie) oraz rozproszonych danych i usługach sieciowych. Przepływy naukowe różni od biznesowych przede wszystkim to, że są zorientowane na wymianę i analizę danych i mogą być bardzo kosztowne obliczeniowo (często wymagają obliczeń równoległych i/lub Gridowych) [1][11][19]. Inna charakterystyka dotyczy bogactwa metadanych i adnotacji wynikającego z używania tych samych danych dla różnych celów i w różnych badaniach naukowych – wymaga to dokładnego określenia kontekstu i pochodzenia danych [2][4]. W końcu, naukowcy są raczej indywidualistami i często sami chcą tworzyć własne "przepływy odkrywania wiedzy", podczas gdy w zastosowaniach biznesowych stosowane są wcześniej określone i zautomatyzowane procedury działań.

Przepływy działań naukowych nie tylko sklejają źródła danych, ale także programy aplikacyjne takie jak narzędzia i pakiety analizy, modelowania i symulacji (sdm.lbl.gov/sdmcenter). W odróżnieniu do integracji wirtualnej, w tym podejściu łączenie zapytań do odległych baz danych i wywołania programów aplikacyjnych odbywa się na dużo niższym i proceduralnym poziomie; w szczególności kolejność działań w łańcuchu operacji jest istotna. (W integracji wirtualnej mediator może reorganizować ustalony plan wykonania zapytania). Przepływy działań mogą włączać usługi oferowane w środowisku Grid, dzięki czemu ułatwiona jest współpraca geograficznie rozproszonych zespołów naukowych z zachowaniem reguł odtwarzania, bezpieczeństwa i prywatności w procesach przetwarzania. Dzięki temu użytkownik ma pełną kontrolę nad przebiegiem procesów integracji.

3. Metody rozwiązywania problemów integracji danych naukowych

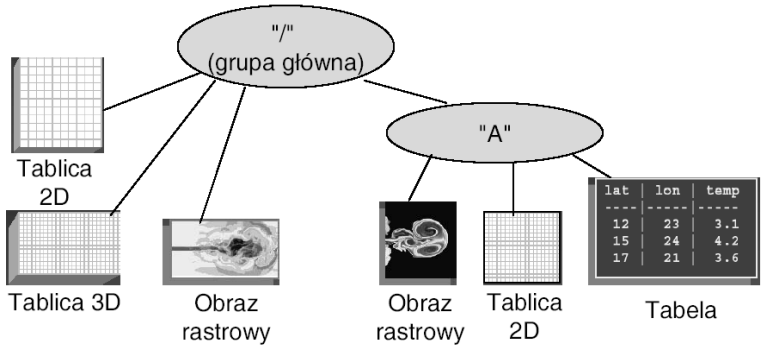
Analizując problemy zarządzania danymi naukowymi można stwierdzić, że w różnych obszarach nauki różne aspekty tego zarządzania mają różną wagę. W przypadku nauk o ziemi mamy do czynienia z ogromnymi zbiorami danych, w przypadku nauk fizycznych dominuje problem obliczeń, a w naukach biologicznych mamy do czynienia z największą różnorodnością struktur danych. Ogromne zbiory danych wielowymiarowych, angażowane w obliczeniach naukowych (na przykład dane geofizyczne uzyskiwane z pomiarów sensorowych) czy technicznych (na przykład metody elementów skończonych), wymagają specjalnych metod organizacji, które w ogóle nie występują w przypadku integracji danych biznesowych.

3.1. Modele danych wielowymiarowych

Przez dane wielowymiarowe rozumiemy przede wszystkim wielowymiarowe tablice. Jako metodę organizacji i przetwarzania tego rodzaju danych przyjmuje się powszechnie standard HDF5 (*Hierarchical Data Format*), na który składa się zarówno format pamiętania struktur złożonych z

grup wielowymiarowych tablic pamiętających dane numeryczne i graficzne, jak również biblioteki oprogramowania realizujące funkcje efektywnego transferu i przetwarzania tych danych (`hdf.ncsa.uiuc.edu/HDF5`). Przykład takiej struktury przedstawia rysunek 1.

Plik HDF5 zaczyna się od grupy głównej (o nazwie "/"). W każdej grupie może być zawartych dowolnie wiele obiektów HDF5, przy czym podstawowymi obiektami HDF5 są grupy i zbiory danych (*datasets*). *Dataset* jest wielowymiarową tablicą elementów danych pamiętaną razem z opisującymi ją metadanymi.



Rys.1. Przykład struktury danych HDF5: grupa główna zawiera trzy zbiory danych (jako tablice wielowymiarowe) oraz jedną podgrupę zawierającą również trzy zbiory danych

Z kolei format XDMF (*eXtensible Data Model and Format*) (`www.arl.hpc.mil/ice`) umożliwia nadanie semantycznego opisu danym HDF5. Powstał on głównie dla potrzeb metody elementów skończonych. XML opisuje metadane, określane jako *dane lekkie* (*light data*), same dane, zwane *danymi ciężkimi* (*heavy data*), pamiętane są w plikach HDF5. Podstawową strukturą sementyczną jest wówczas siatka (ang. *grid*). Każdy *grid* zawiera elementy typów: (a) *Topology* – opisujący elementy siatki i ich przyleganie, (b) *Geometry* – opisujący współrzędne węzłów siatki, (c) *Attribute* – podają wartości w węzłach siatki. Oto przykład pliku XDMF:

```
<XDMF>
  <Domain>
    <Grid Name="Concrete Wall">
      <Topology Type="Hexahedron" NumberOfElements="110520">
        <DataStructure Dimensions="110520 8" Format="HDF"
          DataType="Int" Precision="4">
          NDGM:Wall.h5:/Initial/Connections
        </DataStructure>
      </Topology>
      <Geometry Type="XYZ" >
        <DataStructure
          Dimensions="179685 3" Format="HDF"
          DataType="Float" Precison="8">
          NDGM:Wall.h5:/Initial/XYZ
        </DataStructure>
      </Geometry>
      <Attribute Name="Effective Plastic Strain" Type="Scalar"
        Center="Cell">
        <DataStructure Dimensions="110520" Format="HDF"
          DataType="Float" Precision="8">
          NDGM:Wall.h5:/Results/EffPlaStr
        </DataStructure>
      </Attribute>
    </Grid>
  </Domain>
</XDMF>
```

3.2. Znaczenie adnotacji w integracji danych naukowych

Wiele repozytoriów danych nie posiada ani opublikowanego schematu, ani interfejsu zapytań. Integracja odbywa się poprzez nawigowanie od jednych danych do drugich i z jednego źródła danych do drugiego. Wówczas ważną rolę odgrywają adnotacje. Systemem takim jest DAS (*Distributed Annotation System*) (`www.biodas.org`) działający w architekturze klient-serwer. System umożliwia gromadzenie adnotacji z różnych źródeł internetowych, ich zbieranie i wyświetlanie w jednolity sposób. Wynika stąd waga ontologii i słowników oraz integracji poprzez treść niezależnie od dopasowywania komponentów schematu. Pielęgnacja adnotacji jest zatem ważnym elementem zarządzania danymi naukowymi i ich integracji. Adnotacje są tym samym traktowane jako zasoby wiedzy eksperckiej zrodzone podczas utrzymywania i integracji repozytoriów danych. Adnotacje są propagowane i integrowane wraz z bazami danych 0[2].

3.3. Znaczenie ontologii w integracji danych

Główny problem integracji danych polega nie na różnorodności schematów i modeli danych, ale na luce semantycznej między danymi źródłowymi, która musi być wypełniona przez ekspertów dziedzinowych, aby możliwe stało się łączenie elementu X z jednego źródła z elementem Y z innego źródła. Potrzebne jest "system koordynacji semantycznej", który zapewni mechanizm referencji umożliwiający kojarzenie źródłowych obiektów danych z pojęciami mediatora. Do tego celu służą ontologie, które dostarczają wspólnego słownictwa do wspomagania współdzielenia i wielokrotnego używania wiedzy [1][5][8][14][21]. Można wówczas stosować trójpoziomową architekturę integracji danych:

- źródła danych eksportują nie tylko schematy danych, ale również typy semantyczne reprezentowanych danych, tak aby eksportowane dane były zrozumiałe dla mediatora;
- w schemacie odwzorowań semantycznych reprezentowana jest wiedza o relacjach podobieństwa między dopasowanymi do siebie elementami z różnych źródeł;
- w schemacie mediacji reprezentowana jest wiedza międzyorganizacyjna w postaci globalnych pojęć i ich związków z heterogenicznymi reprezentacjami pojęć w różnych źródłach danych.

Ontologie dziedzinowe mogą wspomagać usługi mediatorów przy formułowaniu i przetwarzaniu zapytań, aby uzyskać poprawną odpowiedź na zapytania wydane względem struktury globalnych pojęć. Korzystając z odpowiednich reguł odwzorowań, zapytania takie mogą być reformułowane na konkretnej terminologii każdego źródła danych angażowanego do zdefiniowania pojęć globalnych, a następnie dane wyszukane mogą być łączone i wyrażane w pojęciach globalnej struktury pojęć zrozumiałej przez użytkownika. W ten sposób uzyskujemy przezroczystość integracji. Dostarcza iluzję pojedynczego języka, jednego modelu danych i jednego źródła.

4. Podsumowanie i wnioski

Integracja danych naukowych staje się ważnym i atrakcyjnym obszarem badawczym łączącym w sobie metody i techniki systemów baz danych, inżynierii wiedzy, obliczeń rozproszonych i nauk doświadczalnych. Dzięki rozwojowi infrastruktury technicznej nauki, głównie w zakresie wysoko wydajnych środowisk rozproszonych typu Grid, staje się możliwe praktyczne realizowanie procesów semantycznej integracji danych o bardzo dużych rozmiarach i bardzo bogatej semantyce. W tym celu prowadzone są na świecie intensywne badania, których celem jest dobre rozpoznanie rzeczywistych potrzeb środowisk naukowych w tym obszarze, a z drugiej – proponowanie prototypowych rozwiązań w zakresie integracji danych naukowych. Przedmiotem tej pracy było przedstawienie najczęściej stosowanych modeli i metod w tym zakresie.

Wiele tez dyskutowanych w pracy opartych jest na doświadczeniach wynikających ze współpracy międzynarodowej w zakresie integracji danych biologicznych [16] oraz z realizowanego projektu dotyczącego semantycznej integracji danych XML w środowisku P2P z wykorzystaniem schematów i ontologii (grant MNiSzW nr 1553/T02/2006/31), a także z doświadczeń nad realizowanym prototypem systemu SixP2P (*Semantic Integration of Xml data in P2P environment*) [18][17][22].

Bibliografia:

[1] Boucelma O., S. Castano, C.A. Goble, i in. Report on the EDBT'02 Panel on Scientific Data Integration. SIGMOD Record 31(4), 2002, pp. 107-112.

[2] Bowers S., B. Ludäscher, A Calculus for Propagating Semantic Annotations Through Scientific Workflow Queries. EDBT Workshops, Lecture Notes in Computer Science, 4254, pp. 712-723

[3] Brzykcy G., J. Bartoszek, T. Pankowski, Schema Mappings and Agents' Actions in P2P Data Integration System, Journal of Universal Computer Science Vol. 14, No. 7, 2008, pp. 1048 – 1060.

[4] Davidson S., i in. Provenance in Scientific Workflow Systems. IEEE Data Eng. Bull. 30(4), 2007, 44-50.

[5] Goczyła K., T. Grabowska, W. Waloszek, M. Zawadzki, The Knowledge Cartography - A New Approach to Reasoning over Description Logics Ontologies, SOFSEM 2006, Lecture Notes in Computer Science 3831, 2006, pp. 293-302.

[6] Haas L., Beauty and the Neast: The Theory and Practice of Information Integration, Database Theory – ICDT 2007, Lecture Notes in Computer Science 4353, Springer 2007, 28–43.

[7] Halevy, A. Y., Rajaraman, A., Ordille, J. J., Data Integration: The Teenage Years, VLDB, 2006, 9–16.

[8] Hess G.N., C. Iochpe, S. Castano, Towards a Geographic Ontology Reference Model for Matching Purposes, GeoInfo 2007, INPE, pp. 35-47.

[9] IBM Information Integration, http://www-306.ibm.com/software/data/integration/

[10] Koloniari, G., Pitoura, E., Peer-to-peer management of XML data: issues and research challenges, SIGMOD Record 34(2), 2005, 6–17.

[11] Ludäscher B., C. A. Goble, Guest editors' introduction to the special section on scientific workflows, SIGMOD Record 34(3), 2005, pp. 3-4.

[12] Microsoft BizTalk Server, http://msdn2.microsoft.com/en-us/library/aa286554.aspx.

[13] Microsoft SQL Server Integration Services (SSIS), SQL Server 2005 Books Online, msdn2.microsoft.com/ en-us/library/ms141026.aspx

[14] Namyoun Choi, I., Han, H.: A Survey on Ontology Mapping, SIGMOD Record 35 (3), 2006, 34–41.

[15] Pankowski T., Cybulka J., Meissner A., XML Schema Mappings in the Presence of Key Constraints and Value Dependencies, Emerging Research Opportunities for Web Data Management, Database Theory ICDT 2007 Workshop EROW 2007, CEUR Workshop Proceedings Vol. 229, pp. 1-15.

[16] Pankowski T., E. Hunt, Data Merging in Life Science Data Integration Systems, Intelligent Information Systems. New Trends in Intelligent Information Processing and Web Mining, Advances in Soft Computing, Springer Verlag, 2005, p. 279-288.

[17] Pankowski T., Query propagation in a P2P data integration system in the presence of schema constraints, 1st International Conference on Data Management in Grid and P2P Systems (DEXA/Globe'2008) , Turin, Italy, Lecture Notes in Computer Science 5187, 2008, pp. 46-57.

[18] Pankowski T., XML data integration in SixP2P – a theoretical framework, EDBT Workshop Data Management in P2P Systems (DAMAP 2008), Nantes, France, ACM International Conference Proceeding Series, 2008, pp. 11–18.

[19] Scientific Data Management Research and Development Group, http://sdm.lbl.gov.

[20] Taylor N.E., Z.G. Ives, Reconciling while tolerating disagreement in collaborative data sharing, SIGMOD Conference, 2006, pp. 13-24.

[21] Udrea O., L. Getoor, R.J. Miller, Leveraging data and structure in ontology integration. SIGMOD Conference, 2007, pp. 449-460.

[22] XQuery 1.0: An XML Query Language. W3C XQuery 1.0: An XML Query Language W3C Proposed Recommendation 2006, http://www.w3.org/TR/xquery/.

[23] Yuan, J., Bahrami, A., Wang, C., Murray, M. O., Hunt, A.: A Semantic Information Integration Tool Suite, VLDB, 2006, 1171–1174.

OBIS – Ocean Biodiversity Information System – nowa era w biologii morza

Jan Marcin Węsławski

Instytut Oceanologii PAN, ul. Powstańców Warszawy 55, 81-712 Sopot

Summary: New tool of extreme importance was recently given for the marine environmental scientists. Ocean Biodiversity Information System has reached its maturity, and proved to be the best existing open data base and reference source for information of marine species distributions worldwide. By now OBIS gathered 13 million georeferenced records about over 80 thousand marine species. The main achievement of OBIS team was to overcame difficulties connected with biological data sharing. The main clients of this data repository are marine protected areas designers and spatial planners. Those interested in faunal changes caused by Climate Warming are having global perspective for monitoring and recognition the distribution ranges.

Niepostrzeżenie, w roku 2007 rozpoczął się nowy etap w badaniach życia w morzu. W tym roku, uruchomiono pierwszą globalną, otwartą dla powszechnego użytkownika bazę danych OBIS (WWW.iobis.org). Przygotowania trwały kilka lat i zabrały wiele czasu kilkudziesięciu osobowej, międzynarodowej grupie inicjatorów i konstruktorów bazy. Jej najważniejszą cechą, odróżniającą od wielu podobnych nieudanych inicjatyw jest sukces w gromadzeniu i udostępnieniu wielkiej ilości dotąd niedostępnych danych o występowaniu zwierząt morskich. W 2008r baza gromadzi dane z 230 rozproszonych banków danych, przedstawiające występowanie 80 tysięcy gatunków z 13 milionów sprawdzonych rekordów (każdy rekord to nazwa gatunku wraz z geograficzną pozycją jego znalezienia) z obszaru całego Oceanu, od wód pływowych do największych głębi.

Żeby zrozumieć znaczenie inicjatywy OBIS trzeba poznać sposób powstawania danych w dziedzinie biologii i ekologii morza. Najważniejszym sukcesem twórców tego systemu było przezwyciężenie naturalnego dla naukowców odruchu chowania i chronienia swoich danych. W przypadku biologów było to wyjątkowo trudne, z powodu specyfiki gromadzonej przez nich informacji. Istniejąca od blisko stu lat skuteczna międzynarodowa wymiana danych meteorologicznych a od kilku lat danych hydrograficznych opiera się na udostępnianiu informacji, którą charakteryzuje niski jednostkowy koszt (jeden pomiar), mała pracochłonność i w ogromny stopień zestandaryzowania. Typowym przykładem jest używanie międzynarodowo skalibrowanych przyrządów do pomiaru temperatury i zasolenia. W czasie jednego zanurzenia sondy uzyskuje się serię kilkuset danych, niemal gotowych do użycia i rozpowszechnienia. Często dane takie dostarczane są przez wyszkolonego technika, a tylko duży ich zbiór jest kontrolowany jakościowo przez naukowca. W przypadku pojedynczego zanurzenia sieci planktonowej – typowego narzędzia biologów morza uzyskujemy tylko materiał przeznaczony do dalszej obróbki. Materiał ten trzeba przetransportować do laboratorium na lądzie, podzielić na wstępne kategorie, które udostępnia się specjalistom, i następnie po kilkunastu godzinach analiz mikroskopowych uzyskuje się informację, którą można wpisać do tabeli Excel i wprowadzić do bazy danych. Informacja ta ma charakter autorski, ponieważ każdy specjalista (taksonom) identyfikuje swoją grupę organizmów wedle najlepszej praktyki i wiedzy, nigdy jednak nie jest to informacja całkowicie obiektywna i standardowa. W praktyce, po zakończeniu morskiego rejsu fizycy (hydrografowie) mają gotowe, zawierające wiele tysięcy danych tabele wyników. Biolodzy, po zakończeniu rejsu przystępują do pracy trwającej zwykle rok, i po tym czasie udostępniają wyniki analiz nie więcej niż setki prób, wymagające krytycznej dyskusji, czyli dane, które mają typowe cechy danych autorskich chronionych przez IPR. Nawet tak pracochłonne, wyniki pojedynczej serii prób biologicznych zwykle nie nadają się do opublikowania – trzeba zebrać ich odpowiednią ilość, nigdy jednak nie będzie ich tyle, by spełniały wymogi statystycznych opracowań typowych dla analiz fizycznych.

Łatwo więc teraz zrozumieć jak wielkim poświęceniem jest oddanie do wspólnego – anonimowego banku danych swojej okupionej wielkim wysiłkiem i opartej na wieloletnim doświadczeniu pracy. Tym bardziej, że po oddaniu do bazy, dane biologiczne zyskują nowy wymiar – pojedyncze serie

wyników dostarczone przez jednego specjalistę znajdują się pośród wielu tysięcy podobnych danych – komfort o jakim żaden biolog nie mógł dotychczas nawet marzyć. Teraz, przyczynkowa, pojedyncza informacja, znajduje się w kontekście zestandaryzowanych, geograficznie zorientowanych podobnych danych. Popularne w XIX i wyśmiewane w drugiej połowie XXw spisy gatunków – tzw. check lists, przez nowe nurty w biologii były zwane „zbieraniem znaczków pocztowych" i odmawiano im prawa do samodzielnej wartości naukowej. Od początku lat 90-tych XXw rozpoczął się trwający do dziś renesans działań inwentaryzacyjnych, taksonomicznych i zoogeograficznych. To, co kiedyś było przyczynkarstwem i zajęciem godnym amatorów, teraz stało się jedną z najważniejszych dróg do poznania kluczowych problemów współczesnego świata- globalnej zmiany klimatu i spadku różnorodności biologicznej. Co ważniejsze wprowadzane do prawodawstwa wszystkich państw morskich zasady ochrony i zrównoważonego gospodarowania w morzu wymagają posiadania aktualnej podstawowej informacji o zasobach żywych – już nie tylko tych eksploatowanych komercyjnie (ryby) ale wszelkich wartościach przyrodniczych środowiska.

Z jednej strony pojawiło się więc zapotrzebowanie – naukowe, społeczne i polityczne, z drugiej strony pojawiło się gotowe narzędzie – OBIS. Przełom w możliwościach zawdzięczamy przede wszystkim systemom gromadzenia i przetwarzania danych – dopiero w wielkiej masie, przyczynkarskie dane stały się poważnym materiałem naukowym, podstawą do coraz liczniejszych publikacji w NATURE i SCIENCE. Pytania, które można teraz stawiać dotyczą globalnych zmian zasięgu, zmian częstości obserwacji poszczególnych gatunków, zależności zanikania populacji od poziomu stresu wytwarzanego przez człowieka, itd. Zestawienie wielkiej ilości danych pozwoliło na pokazanie naszych obszarów niewiedzy. O ile z wód powierzchniowych i płytkich, mamy- w skali globalnej ogromną ilość obserwacji, z dość pełnym pokryciem powierzchni całego wszechoceanu, to już głębokości poniżej 200m, widzimy wiele białych plam, których ilość szybko rośnie z głębokością. Średnia głębokość wszechoceanu to 3500m – czyli większa część powierzchni naszej planety znajduje się pod wodą o tej głębokości. Dane biologiczne z 3500m to zaledwie kilkanaście grup punktów na mapie świata, z zagęszczeniem przy wybrzeżach Europy i USA.

Ambitny i szlachetny cel, który postawiła przez sobą UE i UNESCO to zastopowanie spadku różnorodności biologicznej do 2010r. Cel tyleż nierealny praktycznie co niemożliwy do osiągnięcia nawet w teorii – po prostu dlatego że nie mamy pojęcia jaka jest obecna różnorodności biologiczna. Dotyczy to szczególnie morza, i najdobitniej ilustruje to OBIS. Wyszukane programy statystyczne, które z zebranych danych prognozują ile jeszcze gatunków pozostało do odkrycia dają rozrzut od 1 mln do 20 milionów – w zależności jakie dane z OBIS zostaną użyte. Dlatego kilka wielkich projektów naukowych i kilkuset specjalistów na całym świecie pracuje nad opisaniem ogólnych prawidłowości występowania fauny i flory morskiej, a wciąż publikowane nowe artykuły przynoszą więcej nowych pytań niż odpowiedzi – np. dlaczego obszar Antarktyki jest jednym z najbardziej zróżnicowanych regionów morskich świata, skoro większości regionów, liczba gatunków spada konsekwentnie wraz z obniżeniem temperatury wody. OBIS pozostaje jedynym istniejącym źródłem informacji o występowaniu morskich organizmów w skali globalnej. Problemem dla naukowców pozostaje decyzja czy publikować nowe syntezy na podstawie już dostępnych w OBIS informacji i być pierwszym z nowym pomysłem, czy poczekać jeszcze parę miesięcy na wciąż rosnący zasób informacji i opublikować coś lepiej osadzonego w materiale faktograficznym.

Z podziękowaniami dla Katarzyny Błachowiak Samołyk IOPAN, Mark Costello Leigh Marie Laboratory NZ , Edward Vanden Berge OBIS USA