POLSKIE TOWARZYSTWO INFORMACJI PRZESTRZENNEJ
ROCZNIKI GEOMATYKI 2019 ○ Tom XVII ○ Zeszyt 4(87): 189–206

Artykuł naukowy

Przetwarzanie historycznych rękopisów z wykorzystaniem standardów OGC i bazy danych przestrzennych

OGC standards and spatial database in gathering information from historical sources

Grzegorz Myrda

Instytut Historii Polskiej Akademii Nauk

Abstract

Converting analogue data into digital ones usually is not possible in the case of medieval manuscripts. In turn traditional methods used so far are time-consuming and usually result in the publication containing only effects without the source materials as a foundation. As part of the work carried out at the Institute of History of the Polish Academy of Sciences, tools were developed to improve the work on historical manuscripts (not only maps), using GIS technologies and operating in accordance with OGC standards. Importantly, the role of GIS tools, spatial databases and appropriate standards is not limited to the location of settlements on the map. Their use goes beyond the typical use cases, because they are also used to handle non-geographic data, i.e. the content of manuscripts. As a result, works are progressing faster, are less costly, and it is easier to achieve the synergy effect between projects using repeatedly the same base data, or enriching them adding subsequent information layers. The same tools are also used to publish the results of the work. In connection with the appropriate ontology built for this purpose, as a result we obtain homogeneous and at the same time comprehensive tools for recording, analyzing and publishing changes in settlement occurring in time and space, over hundreds of years of history covering the territories of former and present Poland.

Słowa kluczowe: baza danych przestrzennych, OGC, HGIS, geografia historyczna
Keywords: spatial database, OGC, HGIS, historical geography

Wstęp

Aby stworzyć mapę historyczną, która jest czymś więcej niż tylko obrazem będącym wynikiem skanowania, niezbędne jest przekształcenie szeregu informacji występujących

w materiałach źródłowych, do postaci umożliwiającej dalsze przetwarzanie zawartej tam wiedzy. Im starsze materiały będące podstawą prac, tym częściej są to rękopisy, a więc materiały trudniejsze w odczycie. Docelowa postać danych, powstających w wyniku tworzenia cyfrowej wersji rękopisu, zależy głównie od potrzeb związanych z dalszym wykorzystywaniem tych danych. Bardzo często zdarza się, że opracowane dane są wykorzystywane nie tylko do opracowywania wniosków w ramach projektu macierzystego, ale sięga się po nie także w innych projektach. Potrzebna jest więc jakaś forma bazy danych.

Aby osiągnąć cel, w postaci przejścia od papierowego rękopisu do postaci bazodanowej, rozważenia wymagają przede wszystkim dwa zagadnienia. Po pierwsze, problem zapisu danych w taki sposób aby nie utracić informacji o ich pochodzeniu, czyli o miejscu w manuskrypcie gdzie występowała informacja będąca podstawą ich wprowadzenia do bazy danych. Co ważne, z jednoczesnym zachowaniem cechy łatwego przetwarzania tej referencji, przez już istniejące, standardowe algorytmy i oprogramowanie komputerowe typu COTS[1]. Bez konieczności tworzenia nowych metod, i ich implementacji, obsługujących połączenie między obrazem a wpisem w bazie danych. Drugi problem to model zmian zachodzących zarówno w przestrzeni, jak i w czasie, dla jednostek osadniczych, czyli danych stanowiących fundament i punkt odniesienia dla wielu innych informacji. Wzmianki o nich występują w różnych kontekstach, w różnego typu materiałach źródłowych, pochodzących z różnych okresów czasu. Przy czym najważniejsze cechy miejscowości, stanowiące o ich tożsamości, takie jak nazwa, charakter (typ), a także położenie (Harbelot, 2013), z biegiem czasu zmieniały się. Czy po zmianie jednej lub więcej z takich cech, mamy do czynienia wciąż z tą samą miejscowością? Jakie zmiany muszą nastąpić aby jednostka utraciła swoją tożsamość? Odpowiedzi na tego rodzaju pytania są jednym z wymagań w czasie budowy bazy danych zawierającej historię jednostek osadniczych. Odpowiedzi te musi również uwzględniać zastosowany model danych. Kiedy powstała dana miejscowość? Ile jednostek o określonym charakterze występowało na danym terenie, w zadanym czasie? Które miejscowości zmieniały się najczęściej? Baza danych oparta o opisany tutaj model danych, może pomóc odpowiedzieć na tego rodzaju pytania, pozwalając jednocześnie na analizy przebiegu zmian w czasie i przestrzeni oraz przedstawienie rezultatów na mapie.

[1] COTS (ang. Commercial Off-The-Shelf – prosto z półki). Określenie dla komercyjnej wersji oprogramowania, produkowanej seryjnie i sprzedawanej klientowi bez żadnego dostosowywania do jego potrzeb. Takie oprogramowanie jest tańsze od rozwiązań tworzonych na indywidualne zamówienie. Różnica analogiczna do różnicy pomiędzy odzieżą kupowaną w sklepie, a odzieżą zamawianą u krawca.

Dane wejściowe

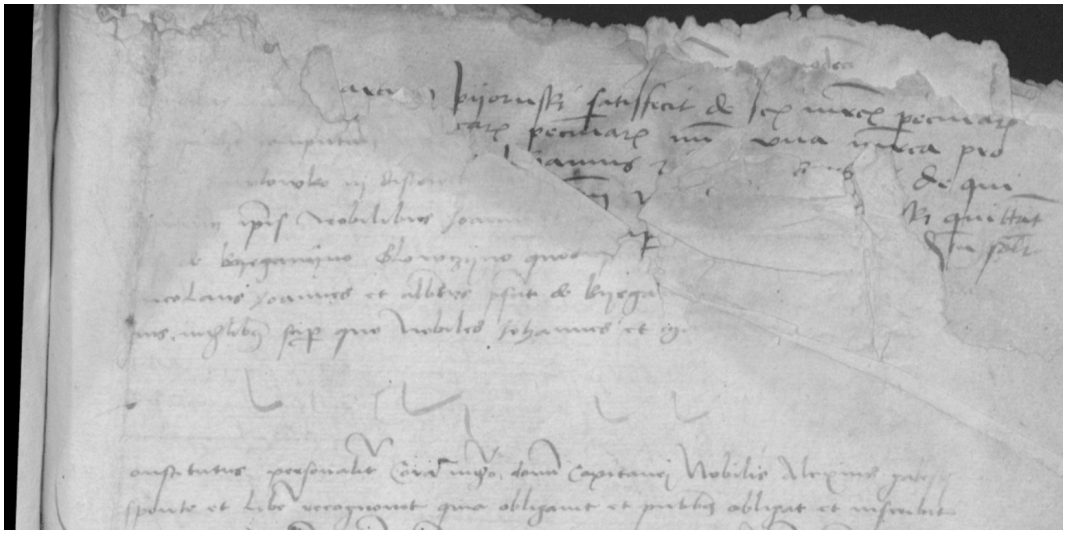
Dane geograficzne z przeszłości to nie tylko dawne mapy, ale pośrednio także część zawartości rękopisów, takich jak księgi sądowe[2] czy rejestry poborowe[3], które zawierają między innymi informacje o ówczesnych miejscowościach i podziale administracyjnym. Ponieważ są to bardzo często obszerne teksty (pojedyncza księga może zawierać nawet kilka tysięcy stron), dlatego warunkiem wejściowym zapewniającym powstanie bazy danych w rozsądnym czasie, jest przetwarzanie ich do postaci cyfrowej w sposób selektywny. Oznacza to, że do bazy danych nie trafia cała treść rękopisu, słowo po słowie, ale wybrana kategoria informacji, w postaci wzmianek o miejscowościach (Słoń, 2017). Z technicznego punktu widzenia możliwe jest wprowadzanie także dodatkowych informacji, takich jak np. informacje o osobach, ale trzeba pamiętać, że im bardziej kompleksowo następuje zbieranie informacji, tym dłużej trwa opracowywanie pojedynczej księgi.

Warto zaznaczyć, iż z kilku powodów, zastosowanie w pełni technologii OCR, która zwykle oszczędza sporo czasu, w tym przypadku (w większości sytuacji) nie jest możliwe. Po pierwsze dlatego, że najczęściej mamy do czynienia z pismem odręcznym, którego styl zdarza się, że ulega zmianie w ramach tego samego rękopisu (zmiana tzw. ręki pisarskiej, czyli osoby wypełniającej np. rejestr poborowy), co może powodować konieczność uruchamiania za każdym razem, czasochłonnego procesu uczenia się przez system OCR nowego charakteru pisma. Po drugie, ze względu na słabą czytelność niektórych materiałów spowodowaną licznymi przebarwieniami, ubytkami papieru, nieuporządkowanymi dopiskami, z którymi komputer sobie nie poradzi. Przykład typowej strony rękopisu przedstawiono na rysunku 1.

Po trzecie, nawet dobrej jakości materiały sprawiają czasem kłopoty w ich odczytaniu, także specjalistom w zakresie paleografii[4]. Po czwarte wreszcie, w trakcie odczytu materiału dokonuje się bardziej jego przekładu (choć nie jest to transkrypcja pełnotekstowa), niż transliteracji, czyli nie oddaje się wiernie li tylko wyglądu tekstu, litera po literze. Nawet gdyby w większym lub mniejszym stopniu wykorzystywać OCR (co w pewnych sytuacjach byłoby możliwe), to nie zastąpi on pracy człowieka, który

[2] Księga sądowa: rejestr zawierający wyroki i postanowienia sądów ziemskich i grodzkich, w tym nazwiska stron i oznaczenia nieruchomości oraz miejscowości.
[3] Rejestr poborowy: wykaz wpłat nadzwyczajnego podatku zwanego poborem przeznaczonego na cele wojskowe w dawnej Polsce. Był swego rodzaju wykazem majątku, zawierającym między innymi spis miejscowości.
[4] Paleografia: nauka pomocnicza historii, zajmująca się badaniem dawnego pisma

na podstawie kontekstu musi dodatkowo nadać dygitalizowanym treściom odpowiednią semantykę.



Rys. 1. Jedna ze słabo czytelnych stron księgi Gr.30 sądu szlacheckiego, dla powiatu kaliskiego z lat 1498–1504

Nie interesuje nas bowiem występujący w tekście sam ciąg znaków Nowa Wieś. Interesuje nas jego znaczenie, czyli jednostka osadnicza o nazwie Nowa Wieś, posiadająca konkretne umiejscowienie w przestrzeni geograficznej i co równie ważne w czasie. Konsekwencją tego jest to, że oprócz nazwy, do bazy danych wprowadzone być muszą co najmniej współrzędne geograficzne oraz data. Można sobie wyobrazić ewentualny udział technologii OCR, użytej na początkowym etapie, do wstępnego przetworzenia obrazu w ciąg liter, w sytuacjach kiedy to byłoby technicznie możliwe. Nie zmieni to jednak faktu, że bez zaangażowania historyków nie jest możliwe stworzenie bazy danych, której zawartość posiada nadane odpowiednie znaczenie, umożliwiające następnie odpytywanie semantyczne. Ewentualne włączenie w cały proces dodatkowego etapu z wykorzystaniem OCR w jego początkowej fazie, mogłoby pomóc w transliteracji, ale w nadawaniu poszczególnym słowom znaczenia już nie.

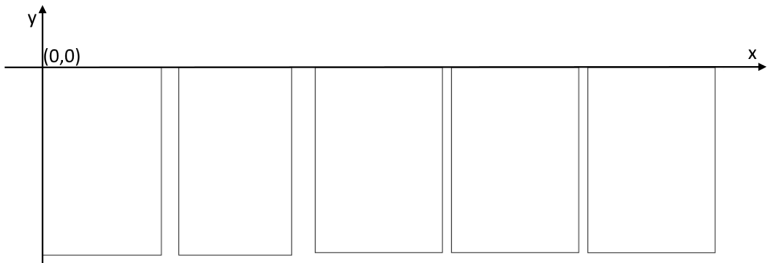
Koncepcja zapisu danych w sposób umożliwiający ich powiązanie z konkretnymi miejscami na skanie rękopisu

Fizycznie danymi wejściowymi całego procesu są zeskanowane poszczególne strony materiału źródłowego. Podstawową warstwą danych są więc dane typu rastrowego, na

które powinny być nakładane kolejne warstwy danych o charakterze wektorowym, gromadzone w postaci bazy danych. Z technicznego punktu widzenia, mamy więc do czynienia z typową sytuacją zarządzania danymi przestrzennymi (choć niegeograficznymi). Gdzie w przestrzeni dwuwymiarowej, znajdują się wszystkie strony rękopisu, po których możemy poruszać się za pomocą współrzędnych w płaskim układzie odniesienia. Skany mogą być rozmieszczone w dowolny sposób, ale dla ułatwienia ich obsługi można przyjąć, że będą ułożone obok siebie w równych odstępach, jak poszczególne klatki na kliszy fotograficznej. Odpowiednikiem tradycyjnego podawania numeru strony i numeru akapitu bądź wiersza na stronie, w celu przejścia do konkretnego zapisu w księdze, jest podawanie współrzędnych w przyjętym dla danej księgi układzie odniesienia. Może nim być dowolny płaski układ współrzędnych, najlepiej zapisany jako metadane dla tego zbioru danych.

Organizacja przestrzeni zawierającej zeskanowane strony

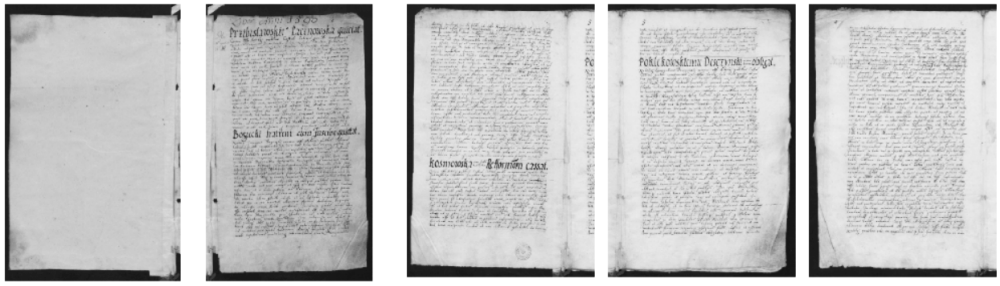
Aby wprowadzać informacje opisowe (czyli dane typu wektorowego), niezbędne jest wcześniejsze odpowiednie rozmieszczenie poszczególnych stron (czyli danych rastrowych) w przestrzeni. W tym celu, dla każdej przetwarzanej księgi, skany wszystkich jej stron agregowane są w jedną całość, posiadającą wspólny układ odniesienia, tak jak to zostało zilustrowane na rysunku 2.



Rys. 2. Układ odniesienia dla typowej księgi, wraz z granicami poszczególnych kart dla księgi sądowej powiatu kaliskiego z roku 1590, określonymi przez referencje przestrzenne

Poszczególnym kartom księgi przypisywane są współrzędne referencyjne w celu ich wpasowania w większą całość (proces analogiczny do przypisywania punktów georeferencyjnych poszczególnym arkuszom mapy rastrowej, w przypadku obszaru geograficznego na który składa się wiele arkuszy map). Karty mogą być rozłożone

w dowolny sposób, zachowując tylko jeden warunek: aby się na siebie nie nakładały. Referencja przestrzenna dla każdej z nich jest definiowana za pomocą pliku WorldFile[5].



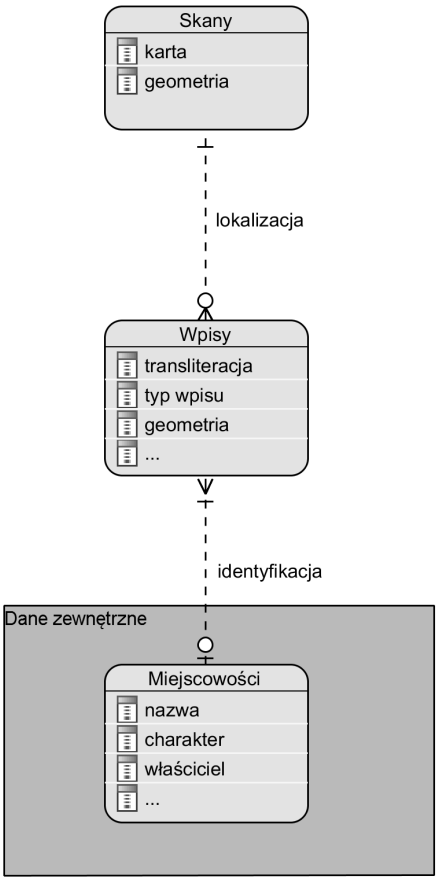
Rys. 3. Karty z księgi sądowej powiatu kaliskiego z roku 1590 rozmieszczone w układzie współrzędnych z rysunku 2

Do każdego zdygitalizowanego wpisu w bazie danych możemy odwoływać się poprzez jego współrzędne, określające jednocześnie jego lokalizację w księdze.

Treść bazodanowa określana jest za pomocą dowolnie kształtowanej bazy danych, której struktura może być dopasowywana do charakteru i zawartości informacyjnej konkretnego rękopisu. Nie istnieje arbitralnie narzucona struktura danych, której należy używać w każdej sytuacji. Informacje wprowadzane są do tabel bazodanowych, o strukturach odpowiadających wybranym pojęciom występującym w danym rękopisie. Każdy typ księgi może posiadać indywidualnie zdefiniowaną strukturę atrybutów opisowych na podobnej zasadzie jak GIS-owa klasa danych. Kluczowe jest natomiast istnienie atrybutu zawierającego współrzędne wpisu, łączące go z odpowiednim miejscem na skanie. Na poziomie technicznym przejawia się to w istnieniu dodatkowej kolumny w tabeli atrybutowej, przechowującej dane typu geometrycznego (pole bazodanowe typu geometry).

Przedstawiony na rysunku 4 ogólny model danych, ilustruje relacje pomiędzy trzema rodzajami danych. Poszczególne wpisy w bazie danych (czyli wzmianki o danej klasie obiektu, na przykład o miejscowościach) opisane są za pomocą dowolnie rozszerzalnej listy atrybutów.

[5] World File: plik tekstowy towarzyszący plikowi z danymi rastrowymi, opisujący położenie obrazu w przestrzeni w zadanym układzie odniesienia.



Rys. 4. Model pojęciowy organizacji danych o miejscowościach występujących w manuskrypcie

Z tego powodu nie istnieje jeden uniwersalny model fizyczny odzwierciedlający powyższy model pojęciowy. Poprzez używanie wspólnego układu współrzędnych, wpisy połączone są przestrzennie z odpowiednimi kartami zeskanowanych materiałów źródłowych. Z kolei opisując każdy wpis odpowiednim identyfikatorem z zewnętrznej bazy danych, można do niego przyłączać dodatkowe zestawy danych. Rolą aplikacji takich jak na przykład wzmiankowany w dalszej części INDXR, jest uwolnienie użytkownika od posługiwania się identyfikatorami, na rzecz używania nazw i mapy. W przypadku pracy nad historyczną mapą punktów osadniczych, rolę referencyjnej, zewnętrznej bazy danych pełni Atlas Historyczny Polski IH PAN, gdzie każda miejscowość, oprócz szeregu charakteryzujących ją atrybutów, posiada swój unikalny identyfikator.

W celu zarządzania połączeniem pomiędzy informacją graficzną (skany), a informacją bazodanową (transkrypcja lub transliteracja) możliwe jest użycie dowolnej aplikacji typu GIS. Za pomocą wrysowania obiektu geometrycznego typu wielokąt (najczęściej wystarcza prostokąt), na skanie strony zaznaczany jest pewien obszar, z którym wiązany jest rekord bazodanowy, podobnie jak w tradycyjnym systemie GIS. Wielokąt może obejmować jeden lub więcej wyrazów i może z nim być związany dowolny zestaw atrybutów opisowych. Abstrahując od zawartości treściowej, z technicznego punktu widzenia, dla tej metody nie ma znaczenia jakie informacje wprowadzamy i jaka jest ich szczegółowość. Im więcej informacji ma być wprowadzanych, tym więcej atrybutów, a być może również powiązanych encji, jest potrzebnych. To, jakie informacje są wprowadzane i jak szczegółowo opisywany jest każdy oznaczony fragment manuskryptu, ustalane jest indywidualnie w zależności od potrzeb konkretnego projektu. W praktyce zależy to zwykle od rodzaju materiału źródłowego i od tego jaki zakres informacji w ramach danego projektu ma być wprowadzany do bazy danych.

Wszystkie wprowadzane w ten sposób do bazy danych wpisy odnoszą się przestrzennie do skanów rękopisów. To oznacza, że każdy wpis posiada referencję do konkretnej strony rękopisu i konkretnego miejsca na tej stronie w postaci lokalizacji geometrycznej. Dlatego uzupełnieniem danych tabelarycznych jest zawsze zbiór skanów poszczególnych stron rękopisu, umieszczony geometrycznie w tej samej przestrzeni co wpisy, umożliwiając tym samym przejście w dowolnym momencie od zapisu w bazie danych do położenia w rękopisie i odwrotnie, wykorzystując do tego celu jedynie współrzędne. Oba zestawy danych (wektorowe wpisy i rastrowe skany) muszą oczywiście używać tego samego układu współrzędnych.

Dla porządku należy wspomnieć, iż alternatywnie do bazy danych, z technicznego punktu widzenia, możliwe jest wykorzystanie tej metody do wprowadzania danych nawet w postaci dobrze wszystkim znanego pliku Shapefile. Celem jest bowiem jedynie zapewnienie możliwości dotarcia z powrotem do miejsca na skanie, które było podstawą wpisu utworzonego w bazie danych, a to zapewnia technologia GIS, niezależnie od formatu zapisu danych. Plik dbf, będący składową formatu Shapefile, można co prawda traktować jak bazę danych, ale bardzo uproszczoną, plikową. Zastosowanie zamiast serwera bazy danych, zapisu bezpośrednio do plików, takich jak Shapefile, otwiera szerzej możliwości wykorzystywania aplikacji typu desktop, szczególnie w niewielkich projektach, gdy większe znaczenie ma niezależność od infrastruktury zewnętrznej (sieć, serwer, aplikacje na serwerze). Opracowane w trybie off-line dane w formacie Shapefile, mogą być na koniec również załadowane do bazy danych, za pomocą standardowych narzędzi GIS, uzyskując ostatecznie taki sam status jak dane wprowadzane od początku, bezpośrednio do bazy danych. Używając zatem tej metody, indeksację materiałów źródłowych można

prowadzić różnymi narzędziami, zarówno w trybie on-line, jak i off-line. Tak opracowane repozytorium danych może być następnie podstawą do publikowania rękopiśmiennych, i nie tylko, zasobów archiwalnych, w trybie materiałów cyfrowych, umożliwiając ich dalsze przetwarzanie. Publikacje mogą być wykonane również za pomocą standardowych narzędzi GIS typu COTS, służących do udostępniania danych o charakterze przestrzennym w Internecie.

W analogiczny sposób możliwe jest wprowadzanie do bazy danych dowolnych klas danych, nie tylko miejscowości, zachowując przy tym połączenie z konkretnym miejscem w źródle (skanie), czyli zapewniając łatwy dostęp do całego kontekstu, bez konieczności wprowadzania całości materiału źródłowego do bazy danych. Opisywana metoda wykorzystana została w praktyce do opracowania wielu różnorodnych treściowo materiałów źródłowych, zarówno pisanych, jak i drukowanych, z których odczytywane były w zależności od potrzeb istniejących w danym projekcie, dane o miejscowościach, osobach, obiektach. Pomimo używania różnych aplikacji i różnych formatów, dane ostatecznie zapisywane były zawsze w tym samym GIS-owym modelu danych.

Aplikacja INDXR

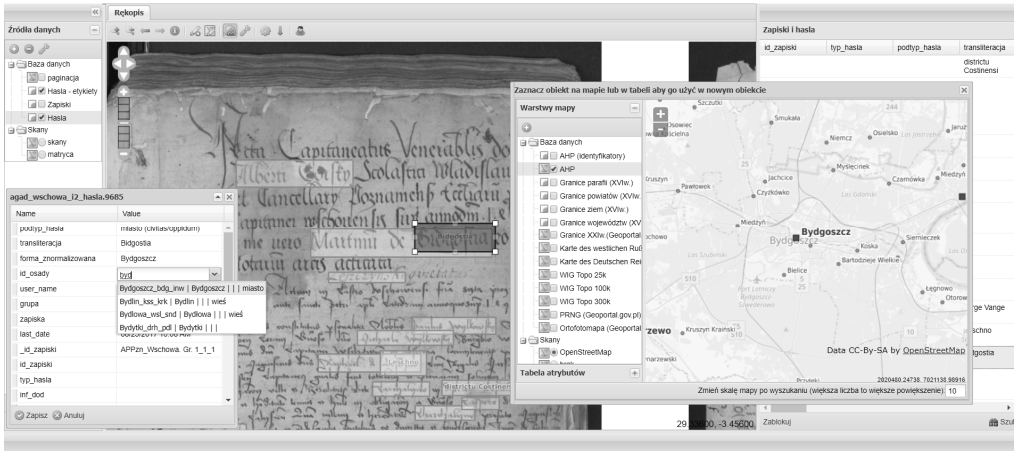
Wykorzystanie uniwersalnego, charakterystycznego dla aplikacji GIS, sposobu zapisu danych, umożliwia pracę z nimi za pomocą dowolnych standardowych aplikacji umożliwiających obsługę danych przestrzennych, w tym także ich internetowych wersji. Nie istnieje zatem aplikacja która jako jedyna umożliwia pracę z danymi opracowanymi w tym modelu. Niemniej jednak w celu zwiększenia wydajności procesu wprowadzania określonych danych, stworzona została specjalistyczna aplikacja internetowa INDXR, ułatwiająca i przyśpieszająca wykonywanie specyficznych operacji na rękopisie. Ponieważ stanowi ona jedną z wielu możliwości wprowadzania danych, należy ją traktować jako pomocniczą aplikację referencyjną. Zbudowana została z gotowych komponentów typu GIS, dzięki czemu nie wymagała angażowania dużych zasobów finansowych, ani czasowych. Zgodnie z popularną zasadą Pareta[6] wypełniła 80% potrzeb za pomocą 20% wysiłku, który byłby potrzebny gdyby budować podobną aplikację od podstaw bez korzystania z technologii GIS.

INDXR działa głównie jako graficzny interfejs użytkownika umożliwiający oznaczanie fragmentów skanu i zapisywanie ich jako geometrii obszarowych w bazie danych. Wykorzystywany jest do tego przestrzenny typ danych występujący we współczesnych systemach zarządzania relacyjnymi bazami danych. Aplikacja, jak każdy typowy system

[6] Zasada Pareta: inaczej zasada 80 na 20, zgodnie z którą 20% badanych obiektów związanych jest z 80% zasobów.

GIS, pozwala na opisywanie obszarów na skanach, za pomocą szeregu atrybutów, określających jakie informacje zawarte są w tym fragmencie rękopisu.

Jest więc typowym, albo w porównaniu z innymi aplikacjami GIS, nietypowym edytorem danych przestrzennych. Na rysunku 5 przedstawiono moment wprowadzania identyfikatora miejscowości, który będzie związany z zakreślonym na rękopisie prostokątem.



Rys. 5. Księga sądowa dla Wschowy (1495–1526) w aplikacji internetowej INDXR

Do standardowych funkcjonalności z kategorii prezentacji danych typu rastrowego w celu edycji danych wektorowych, w ramach aplikacji dodane zostały specyficzne funkcje do automatyzacji często wykonywanych czynności, takich jak na przykład indywidualne listy podpowiedzi, czy mechanizm stronicowania danych rastrowych. Aplikacja INDXR jest z technicznego punktu widzenia specjalizowanym, ale standardowym edytorem GIS. Stosując na przykład QGIS, można tą samą metodą indeksować rękopisy:

- online poprzez usługę WFS[7] (baza danych jest wtedy „ukryta" za serwerem danych przestrzennych, zapewniającym komunikację za pomocą usług zgodnych ze standardami OGC[8]), co jest wskazane jeśli nie chcemy wiązać się z konkretną bazą danych,

[7] WFS: ang. Web Feature Service, usługa zgodna ze standardem OGC o tej samej nazwie pozwalająca na dostęp przez internet do danych w postaci wektorowej; https://www.opengeospatial.org/standards/wfs

[8] OGC: ang. Open Geospatial Consortium, międzynarodowa organizacja niekomercyjna, zajmująca się tworzeniem otwartych standardów dotyczących danych przestrzennych; www.opengeospatial.org

- online poprzez edycję danych bezpośrednio w bazie danych PostgreSQL, wyposażonej w rozszerzenie PostGIS,
- offline poprzez edycję pliku Shapefile. Dane z pliku mogą być następnie załadowane do bazy danych, jeśli miałyby funkcjonować we wspólnym repozytorium z innymi danymi. Wtedy dalsze korzystanie z nich możliwe jest za pomocą dwóch poprzednich metod.

Cała komunikacja pomiędzy aplikacją a bazą danych, odbywa się za pomocą usług zgodnych ze standardami OGC (przede wszystkim WMS[9] i WFS). Usługi te izolują bazę danych w sytuacjach kiedy nie ma potrzeby bezpośredniego dostępu do danych. Znaczna część funkcjonalności aplikacji INDXR opiera się na możliwościach prezentacji danych oferowanych przez usługi WMS (OGC, 2006), WFS (OGC, 2010) i powiązane z nimi standardy. Na przykład wygląd danych wektorowych w przeglądarce, czy optymalizacje wydajnościowe rastrów (np. kafelkowanie), to elementy widoczne w aplikacji INDXR, ale dostarczane do niej z osobnego serwera danych przestrzennych.

Ponieważ różne materiały źródłowe zawierają informacje różnego typu, każdy rękopis może posiadać swoją indywidualną strukturę danych. Aplikacja nie ma wbudowanej sztywnej struktury dla wprowadzanych danych i dynamicznie dostosowuje się do zmian w strukturze danych, wyświetlając formularze odpowiadające wprost strukturze bazy danych.

W ciągu dwóch ostatnich lat, tą metodą, z użyciem głównie aplikacji INDXR, w mniejszym lub większym stopniu przetworzonych zostało ponad 100 różnego rodzaju ksiąg, obejmujących w sumie kilkadziesiąt tysięcy kart. Używany w tym celu generalny model danych, który dobrze odzwierciedla strukturę materiału źródłowego, nie jest jednak rozwiązaniem umożliwiającym łatwą analizę zmian w osadnictwie, w czasie i w przestrzeni oraz bezproblemową prezentację tych danych na mapie. Do tego celu lepiej nadają się dane zagregowane.

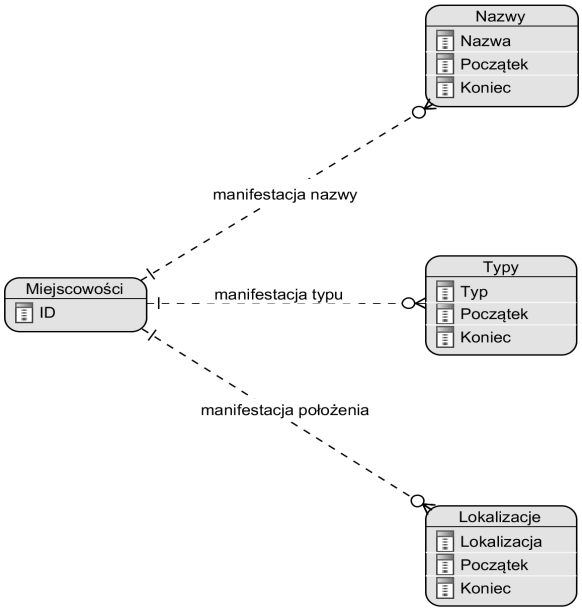
Model danych umożliwiający przechowanie relacji czasoprzestrzennych

Zebrane w przedstawiony powyżej sposób dane, są wykorzystywane w osobnym modelu, zbudowanym w ramach projektu „Ontologiczne podstawy budowy historycznych systemów informacji geograficznej". Model ten został oparty o tak zwane manifestacje, umożliwiające oddanie w bazie danych, stanu jednostek osadniczych, które z upływem czasu ulegały zmianom, głównie w zakresie ich nazewnictwa, typów, położenia, które to cechy stanowią głównie o tożsamości danej jednostki (Garbacz, 2017). Pojęcie

[9] WMS: ang. Web Map Service, usługa zgodna ze standardem OGC umożliwiająca udostępnianie w internecie danych w postaci rastrowej; https://www.opengeospatial.org/standards/wms

manifestacji oznacza tutaj przejaw, uzewnętrznienie pewnych cech. Na przykład jeśli miejscowość w pewnym okresie czasu posiada określoną nazwę, to jest to manifestowane (uzewnętrzniane) przez używanie tej nazwy w różnego rodzaju dokumentach. Relacja pomiędzy konkretną nazwą, a konkretną miejscowością, to manifestacja nazwy dla tej miejscowości, w zadanym okresie czasu. Po zmianie nazwy miejscowości, mamy do czynienia z odrębną manifestacją, ponieważ wartość atrybutu jest już inna. Manifestacja ta może dotyczyć tej samej miejscowości, bądź już innej miejscowości, w zależności od tego czy w wyniku tej i być może innych zmian, nastąpiła zmiana tożsamości. W ten sposób każda cecha obiektu, która nie zmienia się w określonym czasie, jest jego pojedynczą manifestacją. Model pojęciowy zawierający podstawowe cechy został przedstawiony na rysunku 6.

W modelu przedstawionym na rysunku 6, każda cecha jednostki osadniczej stanowi osobną manifestację czasową, z przypisaną wartością tej cechy oraz czasem trwania. Wszystkie manifestacje danej jednostki połączone są ze sobą za pomocą identyfikatora reprezentującego tożsamość danej jednostki. W szczególności położenie jednostki to tylko jedna z cech wyrażona za pomocą manifestacji. Wartością tej manifestacji są współrzędne geograficzne. Ewentualne kolejne cechy miejscowości mogą być wyrażane w analogiczny sposób, za pomocą oddzielnych manifestacji, posiadających wartość cechy, datę początkową, datę końcową, oraz powiązanie z identyfikatorem miejscowości.

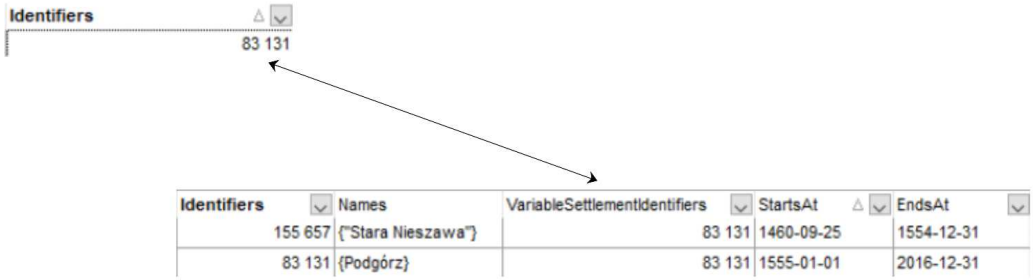


Rys. 6. Schemat pojęciowy modelu czasowo-przestrzennego dla nazwy, typu i lokalizacji miejscowości

Gromadzone w ten sposób informacje są podstawą do opracowywania jednolitej struktury danych o miejscowościach, utożsamiania ze sobą różnych typów miejscowości w różnych czasach i na różnych terenach (na przykład pod różnymi zaborami), a przede wszystkim nadawania tej samej tożsamości konkretnym jednostkom osadniczym. Często, występujące w różnych materiałach pisanych wzmiankowane miejscowości, pomimo tego że różnią się niektórymi cechami (czasami nazwą, czasami typem, czasami lokalizacją), mogą być utożsamiane z tą samą jednostką osadniczą, jeżeli większość cech, albo najważniejsze z nich zostaną zachowane (Szady, Ławrynowicz, 2017; Garbacz, 2018).

Przykład

Powiedzmy, że chcemy zapisać w bazie danych informacje o jednostce osadniczej której wszystkie cechy ulegały z biegiem czasu zmianom. Jedyny niezmienny element to identyfikator tej jednostki. W naszym przykładzie będzie to liczba 83131. Na rysunku 7 widzimy że miejscowość ta w latach 1460-1554 funkcjonowała pod nazwą Stara Nieszawa. Natomiast w latach 1555-2016 jako Podgórz.



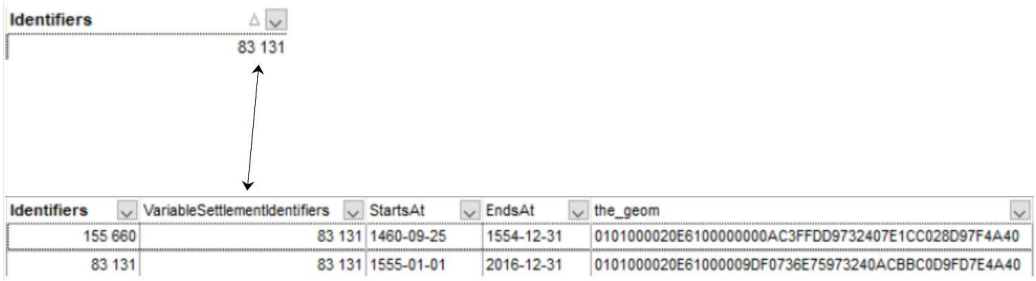
Rys.7. Fragment bazy danych zawierający pojedynczą jednostkę osadniczą o identyfikatorze 83131 oraz powiązane z nią manifestacje nazwy. Współcześnie jest to dzielnica Torunia: Podgórz

Na rysunku 8 widać że cztery razy zmieniał się charakter miejscowości, który określony jest za pomocą kodu typu jednostki osadniczej. W tym przypadku, kody te to: 2- wieś, 3-miasto, 61-część miasta.

Identifiers	VariableSettlementIdentifiers	SettlementTypeIdentifiers	StartsAt	EndsAt
155 657	83 131	2	1460-09-25	1611-11-06
155 656	83 131	3	1611-11-07	1833-03-26
155 655	83 131	2	1833-03-27	1924-12-31
155 654	83 131	3	1925-01-01	1938-03-31
83 131	83 131	61	1938-04-01	2016-12-31

Rys.8. Jednostka osadnicza o identyfikatorze 83131 oraz powiązane z nią manifestacje typu

Z kolei z tabel na rysunku 9 wynika że miejscowość istniała w dwóch lokalizacjach. Współrzędne w kolumnie *the_geom* zapisane są w postaci binarnej. Po dokonaniu rzutowania na typ WKT[10] są to współrzędne odpowiednio: POINT(18.5932311869783 52.9988146722835) i POINT(18.5916356118256 52.9921219054298).

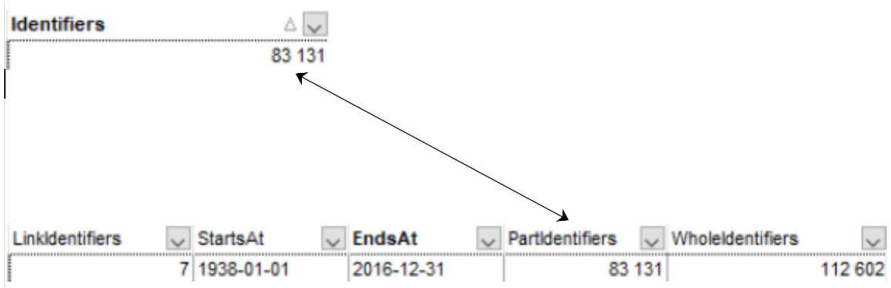


Identifiers	VariableSettlementIdentifiers	StartsAt	EndsAt	the_geom
155 660	83 131	1460-09-25	1554-12-31	0101000020E6100000000AC3FFDD9732407E1CC028D97F4A40
83 131	83 131	1555-01-01	2016-12-31	0101000020E61000009DF0736E75973240ACBBC0D9FD7E4A40

Rys. 9. Jednostka osadnicza o identyfikatorze 83131 oraz powiązane z nią manifestacje lokalizacji

Na rysunku 10 zobrazowano tabele bazy danych o zawartości świadczącej o tym, że od 1938 roku miejscowość ta jest częścią innej miejscowości.

[10] WKT: ang. Well-known text, format zapisu współrzędnych w postaci tekstowej, zgodnie ze specyfikacją OGC Simple Feature Access



Rys. 10. Jednostka osadnicza o identyfikatorze 83131 oraz powiązana z nią manifestacja mereologiczna, reprezentujące przynależność Podgórza (83131) do Torunia (112602), począwszy od 1938 roku

Baza danych przestrzennych

Od dłuższego już czasu producenci baz danych wyposażają je w rozszerzenia przestrzenne, dzięki czemu możliwe jest korzystanie z dodatkowych, geometrycznych typów danych. Rolę serwera bazy danych stanowiącego repozytorium dla wszystkich opisywanych tutaj danych, pełni PostgreSQL, wyposażony w rozszerzenie przestrzenne PostGIS. W przypadku konieczności przetwarzania, a w szczególności wizualizacji, dużych ilości danych o charakterze przestrzennym, wymagana jest duża wydajność, z którą relacyjne bazy danych w porównaniu z bazami grafowymi wciąż radzą sobie lepiej (Garbis, 2013). Wszystkie zbierane informacje, zarówno w modelu wykorzystywanym do indeksacji, opisanym w pierwszej części, jak i w modelu danych umożliwiającym przechowywanie relacji czasoprzestrzennych, rejestrowane są w tej samej bazie danych. Funkcje PostGIS, z jednej strony zapewniają standardowym aplikacjom zewnętrznym dostęp do danych przestrzennych, natomiast z drugiej strony pełnią rolę API[11] najniższego poziomu za pomocą którego możliwe są różnego rodzaju bezpośrednie manipulacje na zindeksowanych treściach skanów, z wykorzystaniem standardowego języka SQL, choć rozszerzonego o funkcje geometryczne. Niezależnie od tego za pomocą jakiej aplikacji rękopisy zostały przetworzone do postaci bazodanowej, można je dodatkowo przetwarzać na tym właśnie poziomie. Umożliwia to realizację funkcji specjalnych takich jak na przykład automatyczne przypisanie słowa lub zwrotu do rozdziału, z wykorzystaniem funkcji ST_Intersects, czyli wyszukania części wspólnej.

[11] API: Programistyczny interfejs dostępowy oferowany przez systemy komputerowe, umożliwiający wywoływanie wewnętrznych funkcji tego systemu.

Podsumowanie

Uniwersalność GIS-owego, ogólnego modelu danych, pozwala na użycie go również do indeksacji rękopisów, zapewniając możliwość połączenia rekordu w bazie danych, z konkretnym miejscem w rękopisie. Dzięki użyciu przestrzeni jako elementu integrującego i zapożyczeniom technologicznym ze świata GIS w trakcie indeksacji rękopisów, nie jest konieczne opracowywanie nowej metody łączenia informacji ze skanów z bazą danych. Nie jest również konieczne kosztowne tworzenie całkowicie nowego oprogramowania. Jednocześnie dzięki stosowaniu standardów OGC w komunikacji zarówno wewnętrznej, jak i zewnętrznej, zachowana zostaje swoboda wyboru narzędzi i technologii wykorzystywanych do dalszych prac.

Z kolei stworzony w kolejnym etapie czasowo-przestrzenny model danych dla zmian w osadnictwie na terenie Polski pozwala na usystematyzowanie danych pochodzących z różnych źródeł i z różnych okresów czasu, umożliwiając budowę historycznego systemu informacji geograficznej (*ang. HGIS - Historical GIS*).

Łącząc dane z obu modeli dysponujemy zarówno materiałem źródłowym w postaci cyfrowej, jak i edycją krytyczną tych danych. Wspólnym rezultatem jest spójny model zmian w osadnictwie, zachodzących w czasie i przestrzeni, przez setki lat historii terenów dawnej i obecnej Polski.

Podziękowania

Dziękuję panom prof. Bogumiłowi Szademu i prof. Markowi Słoniowi za możliwość udziału w realizacji projektów „Ontologiczne podstawy budowy historycznych systemów informacji geograficznej" i „Atlas historyczny Polski XVI wieku - dopełnienie serii" dzięki którym mógł powstać niniejszy artykuł.

Finansowanie

Prace opisywane w artykule zostały zrealizowane w ramach projektów „Ontologiczne podstawy budowy historycznych systemów informacji geograficznej" oraz „Atlas historyczny Polski XVI wieku - dopełnienie serii" finansowanych w latach 2015-2020 ze środków NPRH.

Literatura (References)

Garbacz P., Trypuz R., 2017: Representation of tensed relations in OWL. A survey of design patterns. *Research Conference on Metadata and Semantics Research*: 62–73, Springer.

Garbacz P., Ławrynowicz A., Szady B., 2018: Identity criteria for localities. *Frontiers in Artificial Intelligence and Applications. Volume 306: Formal Ontology in Information Systems*: 47-54, 10.3233/978-1-61499-910-2-47

Garbis G., Kyzirakos K., Koubarakis M., 2013: Geographica: A Benchmark for Geospatial RDF Stores (Long Version). In: Alani H. et al. (eds) *The Semantic Web – ISWC 2013*. *Lecture Notes in Computer Science*, vol 8219, 343-359. Springer, Berlin, Heidelberg.

Harbelot B., Arenas H., Cruz C., 2013: Continuum: A spatiotemporal data model to represent and qualify filiation relationships. *4th ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS) 2013*, Oct 2013, Orlando, United States.

Słoń M., Słomski M., 2017: Edycje cyfrowe źródeł historycznych (Digital editions of historical sources). W: *Jak wydawać teksty dawne, Staropolskie Spotkania Językoznawcze 2*: 65–84, Poznań.

Szady B., Ławrynowicz A., 2018: Considering Identification of Locality in Time: Theoretical and Practical Approach. In: Fogliaroni P., Ballatore A., Clementini E. (eds) *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory*. COSIT 2017. Lecture Notes in Geoinformation and Cartography, 283-291, Springer, Cham.

Open Geospatial Consortium, 2006: Web Map Server Implementation Specification https://www.opengeospatial.org/standards/wms

Open Geospatial Consortium, 2010: Web Feature Service 2.0 Interface Standard https://www.opengeospatial.org/standards/wfs

Streszczenie

Przekształcanie danych analogowych w cyfrowe, w przypadku średniowiecznych rękopisów najczęściej nie jest możliwe w sposób automatyczny. Z kolei tradycyjne, stosowane dotąd powszechnie metody są czasochłonne i kończą się zwykle jedynie publikacją samych wyników prac, bez materiałów źródłowych będących ich podstawą. W ramach prac prowadzonych w Instytucie Historii PAN, powstały metody i narzędzia usprawniające prowadzenie prac nad historycznymi rękopisami (nie tylko mapami), wykorzystujące technologie GIS i działające zgodnie ze standardami OGC. Co istotne, rola narzędzi GIS, przestrzennych baz danych i odpowiednich standardów nie ogranicza się jedynie do zwykłej lokalizacji obiektów przestrzennych na mapie. Ich zastosowanie wykracza poza typowe przypadki użycia, wykorzystywane są one bowiem również do obsługi danych niegeograficznych, choć w pewnym sensie przestrzennych, czyli treści rękopisów. Dzięki temu prace postępują szybciej, są mniej kosztowne, a także łatwiej jest uzyskać efekt synergii międzyprojektowej wykorzystując wielokrotnie te same dane, lub wzbogacając je w kolejne warstwy informacyjne. Co ważne te same narzędzia wykorzystywane są również do publikacji rezultatów prac. W połączeniu z budowaną do tego celu odpowiednią ontologią otrzymujemy w wyniku jednorodne i jednocześnie wszechstronne narzędzia do rejestrowania, analizowania i publikowania zmian w osadnictwie, zachodzących w czasie i przestrzeni, przez setki lat historii terenów dawnej i obecnej Polski.

Dane autorów / Authors details:

mgr Grzegorz Myrda
ORCID 0000-0002-2756-8654
gmyrda@ihpan.edu.pl

Przesłano / Received	10.01.2019
Zaakceptowano / Accepted	25.11.2019
Opublikowano / Published	12.12.2019