Infrastruktura Dariah.lab w Szkole Nauk o Języku i Literaturze UAM

Katarzyna Klessa, Anna Jelec, Maciej Karpiński, Mirosław Wobalis

Fundusze Europejskie Inteligentny Rozwój
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

- Projekt realizowany w obszarze **Rozwój nowoczesnej infr astruktury badawczej sektora nauki**,wpisanej na Polską Mapę Infrastruktury Badawczej.
- Projekt realizowany jest przez **16 Partnerów**, z czego 15 należy do konsorcjum DARIAH-PL *(Digital Research Infrastructure for the Arts and Humanities).* Liderem jest PCSS.
- Sześć modułów infrastruktury w UAM, trzy z nich, realizowane w SNJL UAM prezentuemy Państwu dzisiaj.

Uniwersytet Warszawski

Instytut Badań Literackich PAN

Poznańskie Centrum Superkomputerowo-Sieciowe

Akademia Sztuk Pięknych w Warszawie

Instytut Historii PAN

Instytut Podstaw Informatyki PAN

Instytut Slawistyki PAN

Instytut Języka Polskiego PAN

Politechnika Poznańska

Uniwersytet Adama Mickiewicza

Uniwersytet Jagieloński w Krakowie

Politechnika Wrocławska

Uniwersytet Marii Curie-Skłodowskiej

Uniwersytet Wrocławski

Instytut Sztuki PAN

Uniwersytet Mikołaja Kopernika w Toruniu

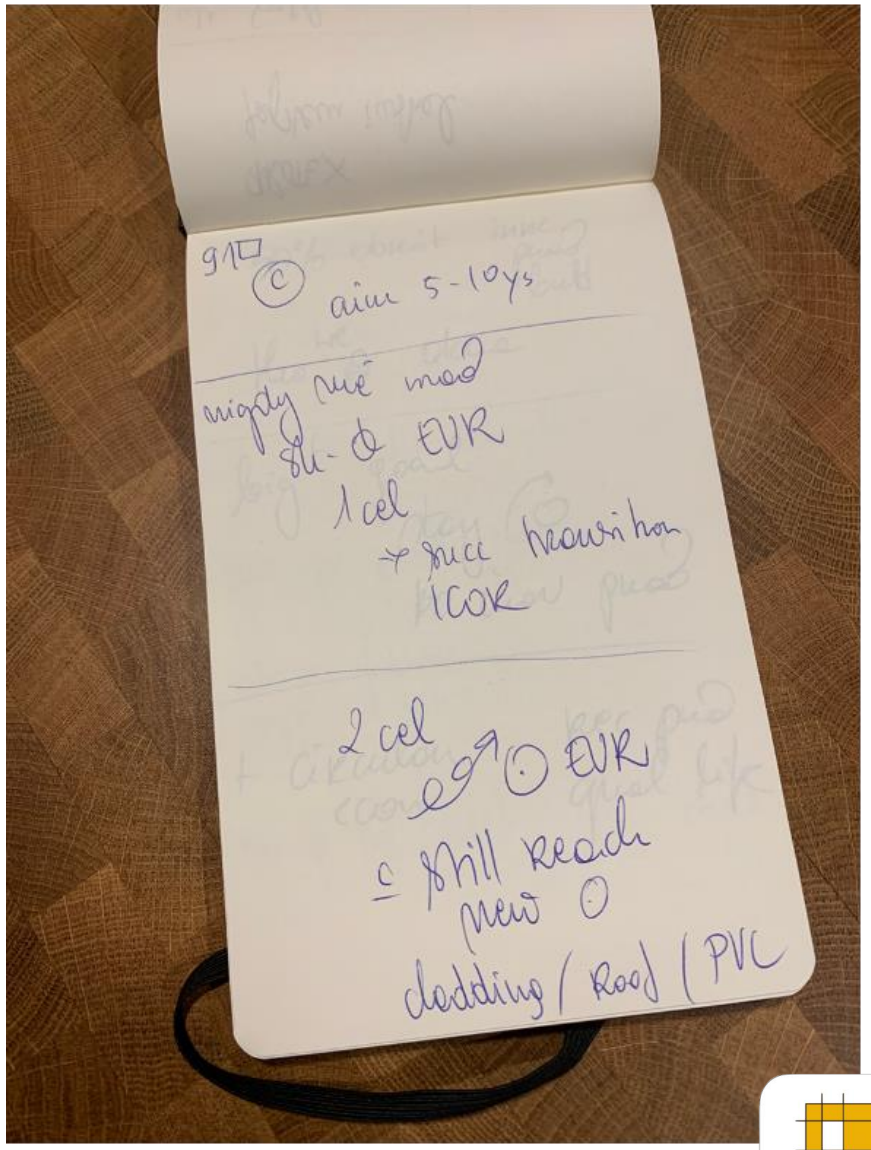
HANOI

Czym jest HANOI

- HANOI, czyli Handwritten Notation of Interpreters to korpus skanów notatek zebranych od zawodowych tłumaczy ustnych oraz studentów i studentek tłumaczenia.
- Za pomocą HANOI możemy: **przeszukiwać** skany notatek zebrane od pracujących aktywnie tłumaczy i tłumaczek; **przeprowadzić ich analizę wizualną**, przeszukać korpus aby **znaleźć notatki** o konkretnych parametrach, **pobrać korpus** w całości oraz **załadować własne notatki**.
- HANOI to również narzędzie wspierające nauczanie tłumaczenia konsekutywnego.
- Załadowanie własnych odręcznie pisanych notatek tłumaczeniowych pozwoli nam odczytać ich parametry oraz porównać jakość naszych notatek z notacją doświadczonych profesjonalistów.

Notacja w tłumaczeniu ustnym

- Tłumaczenie **konsekutywne**: prelegent mówi 3-20 minut, tłumacz notuje a potem tłumaczy na podstawie notacji
- Notacja w tłumaczeniu konsekutywnym stanowi wyjątkowy rodzaj tekstu, bowiem zawsze tworzona jest na styku dwóch języków - tłumacz słuchając wypowiedzi w języku źródłowym notuje wybrane treści w celu odtworzenia ich w języku docelowym.
- Ręcznie robione notatki to jedno z ważniejszych narzędzi w warsztacie tłumacza konferencyjnego.

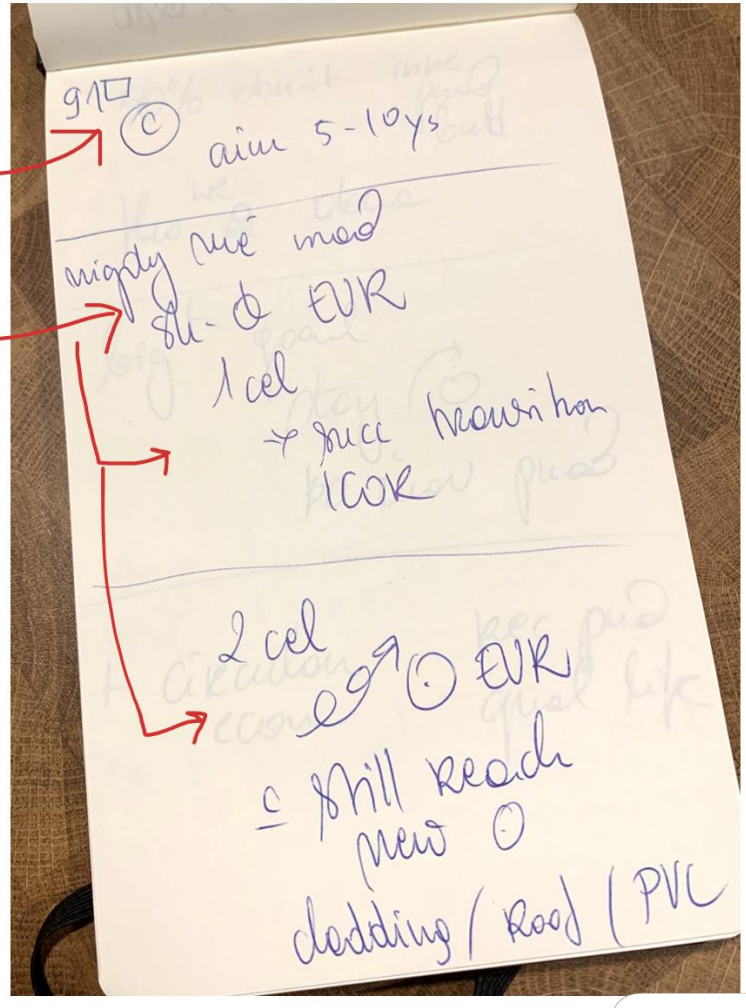


Dlaczego warto badać notatki tłumaczy ustnych?

- Wyjątkowa forma ręcznie pisanego tekstu, różniąca się od wszystkich innych form zapisu, w tym „zwykłych” notatek
- Umiejętności notacji uczy się we wszystkich szkołach tłumaczenia; oparta jest na mnemotechnikach, symbolach, wykorzystaniu kartki i treningu pamięci
- Każdy tłumacz z czasem wypracowuje własny system notacji, możliwy do odczytania tylko dla niej/niego i zazwyczaj jedynie przez krótki czas po realizacji tłumaczenia
- Notatki tłumaczy zawodowych rzadko są zrozumiałe dla kogoś oprócz twórcy
- Mało zbadany proces i produkt tłumaczenia

Cel firmy:
91 krajów
w ciągu 5-10 lat

W Europie z czasem powinno się osiągnąć cele:
1) przejście na system ICOR
2) rozwój rynku + wejście na nowe rynki pokryć dachowych z PVC

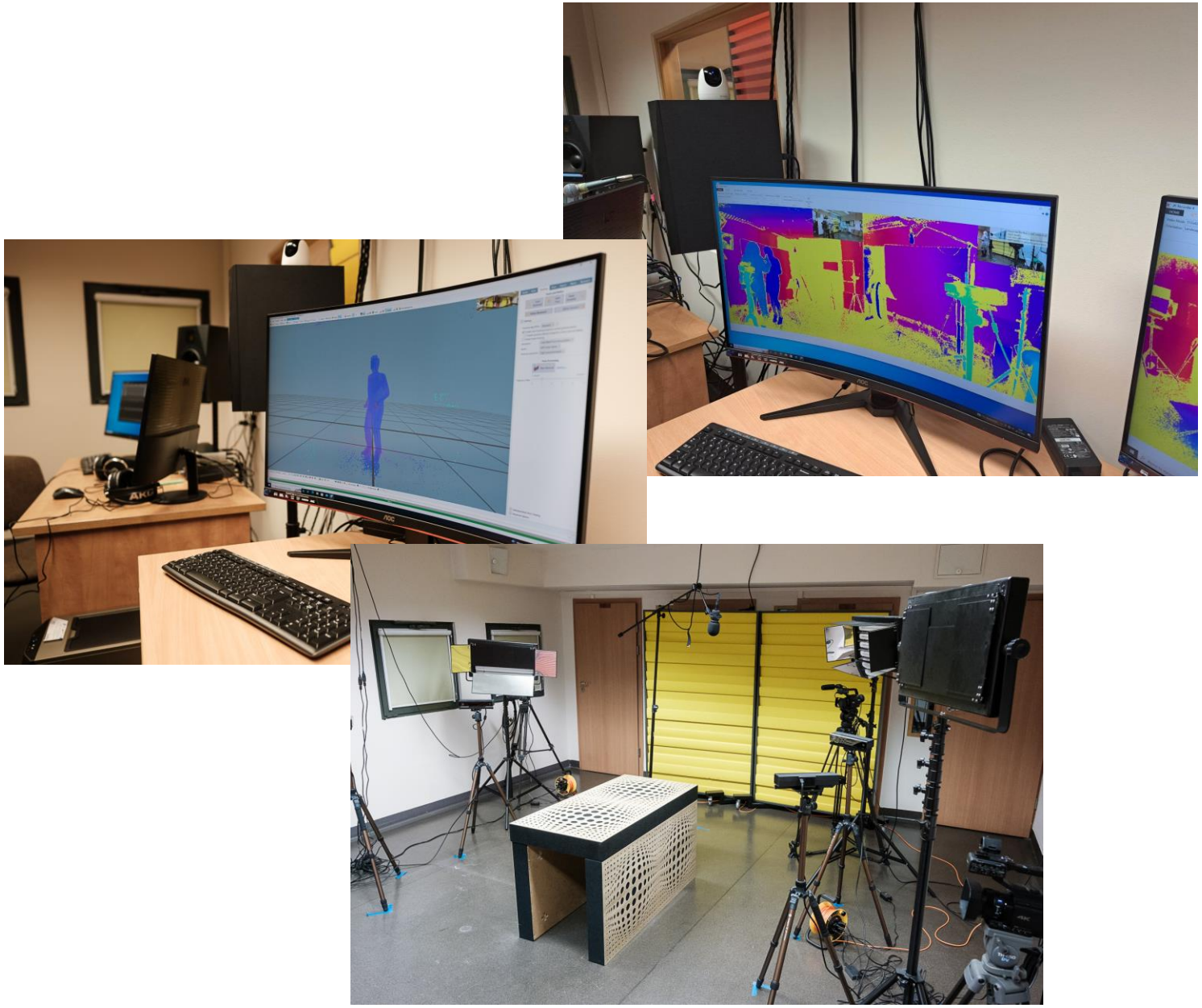


Cel modułu HANOI

- Notatki tłumaczy ustnych to źródło informacji o procesach językowych i poznawczych zachodzących na styku dwóch języków i kultur. Rośnie zainteresowanie tematem tłumaczenia ustnego oraz zachodzących podczas niego procesów. Jednocześnie stosunkowo niewielka dostępność danych sprawia, że wiele hipotez badawczych przetestowano na niewielkich grupach.
- Korpus notacji powstałej w tłumaczeniu ustnym jest pierwszym tego typu projektem. Badania nad tłumaczeniem ustnym obecne od kilkunastu lat w światowej nauce mogą opierać się zarówno na metodach tradycyjnie psychologicznych, wykorzystujących metody obliczeniowe, czy neuropsychologicznych. HANOI udostępnia naukowcom zebrane zasoby do badań empirycznych z udziałem tłumaczy ustnych, których kompetencje zawodowe ukształtowane zostały w środowisku polskim, natomiast adepci i adeptki tłumaczenia skorzystać mogą z możliwości porównania własnych notatek z notacją zawodowców.
- Zastosowanie bazy danych w gospodarce możliwe jest poprzez komercjalizację narzędzia HANOI. Ze względu na charakterystykę ręcznie robionych notatek, digitalizacja tych materiałów może na dłuższą metę posłużyć do usprawnienia narzędzi AI służących do inteligentnego rozpoznawania pisma ręcznego. Korpus ręcznie robionych notatek zawiera będzie bowiem przykłady wielu różnych stylów pisma, symbole, ryciny, skróty i inne sposoby zapisu stając się bezcennym zasobem dla zaawansowanych algorytmów uczących.

HANOI

MultiCo



Multimodalność komunikacji

- *Multimodalność komunikacji* - wykorzystanie w komunikacji różnych **modalności zmysłowych**, m.in. słuchowej i wzrokowej. W obrębie danej modalności można wyróżnić szereg **kanałów** przekazu.
- *Prozodia i gestykulacja nie są dodatkiem* do wypowiedzi, stanowią jej integralną składową.
- Lingwiści powinni posiadać motywację i kompetencję, by analizować wypowiedzi *holistycznie*.

  - (Liebal i in. 2013, Bonacchi i Karpiński, 2014, Jarmołowicz-Nowikow i Karpiński 2020)

Infrastruktura laboratorium do badań nad komunikacją multimodalną

- Zespół: Ewa Jarmołowicz-Nowikow, Maciej Karpiński (koordynator), Katarzyna Klessa, Brygida Sawicka-Stępińska, Janusz Taborek, Michał Piosik (wsparcie techniczne), Elwira Wilczyńska (wsparcie administracyjne), oraz współpracownicy i studenci
- 4 standardowe kamery 4k z możliwością rejestracji oraz oprogramowanie do przechwytywania i edycji wideo
- 4 kamery z sensorami głębi i oprogramowanie do rejestracji, parametryzacji i biomechanicznej analizy ruchu
- System rejestracji i przetwarzania dźwięku, rejestrujący na odrębnych ścieżkach sygnał z maksymalnie ośmiu mikrofonów jednocześnie, wraz z kilkoma typami mikrofonów
- Serwer z oprogramowaniem umożliwiającym bezpieczne gromadzenie, przechowywanie, i udostępnianie danych
- Dysponujemy również aparaturą z wcześniejszych projektów

Korpus MultiCo

- Złożony z trzech głównych podkorpusów, każdy z nich liczący około pięciu godzin nagrań:
  - Monologi: wystąpienia sejmowe oraz mowy motywacyjne z forum TEDx
  - Dialogi: dialogi zadaniowe, zarejestrowane w warunkach laboratoryjnych
  - Multilogi: programy sportowe, poświęcone piłce nożne, rejestrowane w studiu telewizyjnym (TV Sport)

- Synchroniczna anotacja korpusu
  - Transkrypcja ortograficzna i fonematyczna
  - Segmentacja na poziomie jednostek międzypauzalnych, wyrazów, sylab i głosek
  - Manualna anotacja fraz gestowych i wybranych typów ruchów głowy lub tułowia (zależnie od subkorpusu)
  - Rejestracja parametrów ruchu z wykorzystaniem systemu mocap (subkorpus dialogowy)

Zastosowania korpusu i możliwości badawcze laboratorium

- Korpus stanowi wizytówkę laboratorium, pokazując jego niektóre jego możliwości. Będzie pierwszym w Polsce i jednym z nielicznych na świecie dostępnych publicznie i darmowo (do celów badawczych) korpusów multimodalnych.
- Zastosowania korpusu: badania nad multimodalnymi aspektami mechanizmów perswazji i ekspresji i interakcji dialogowej, zasoby do szkoleń w zakresie perswazji i ekspresji, zasoby do tworzenie systemów dialogowych z animowanymi awatarami, i wiele innych.
- Możliwości techniczne laboratorium: rejestracja wizji i fonii w warunkach laboratoryjnych i terenowych, anotacja, analiza i przetwarzanie akustyczna sygnału mowy, anotacja zapisu wideo, analiza danych *motion capture.*

(Część zespołu MultiCO działa również w laboratorium koordynowanym przez prof. Piotra Podlipniaka (IM WNoS) w zakresie badania związków między muzyką, mową, i ruchem.)

Węzeł Filologiczny

Węzeł Filologiczny to zbiór narzędzi informatycznych służących badaniu (a także projektowaniu) tekstów literackich oraz prowadzeniu badań językoznawczych.

W ramach działań konstytuujących cele Węzła umieszczono:

1. Pozyskanie korpusu źródeł służących zaprojektowaniu i przetestowaniu narzędzi informatycznych.

2. Opracowanie i wdrożenie zestawu narzędzi informatycznych służących badaniom językoznawczym i literaturoznawczym.

Dominas / Hojdis / Liberek / Pisarski / Winiecka / Wobalis

Korpus źródeł Węzła Filologicznego

W ramach realizacji Węzła Filologicznego pozyskano lub odzyskano:

- **2288 numerów** (o średniej liczbie stron 4, co daje ogółem około 9152 plików graficznych) ukazującego się w XIX wieku w zaborze rosyjskim „Kaliszanina. Gazety miasta Kalisza i jego okolic” z lat 1870-1892. **Zadanie zamknięte** - archiwum dostępne jest publicznie w zasobach Wielkopolskiej Biblioteki Cyfrowej. Korpus pozyskano od Rosyjskiej Biblioteki Narodowej w Petersburgu.
- **75 tysięcy jednostek** (dokumentów do skanowania) pochodzących z archiwum prof. Stanisława Bąby. **Zadanie zrealizowane w 70%** - zakończono digitalizację i aktualnie
- tworzony jest cyfrowy "Korpus Frazeologiczny Języka Polskiego".
- **990 afiszy teatralnych** z lat 1887-1935. **Zadanie zrealizowane w 85%** - zakończono konserwację, aktualnie trwa profesjonalna digitalizacja afiszy.

Zgodnie z założeniami projektu wszystkie źródła znajdą się (dla celów badawczych) w otwartym dostępie.

Zbiór narzędzi cyfrowych Węzła Filologicznego

Biurko Badacza

Opracowanie kompleksowego systemu analizy źródeł językoznawczych i literaturoznawczych w formie zbioru narzędzi modułowych do wstępnego i kompleksowego przygotowywania materiałów źródłowych (głównie dawnych) o charakterze statycznego tekstu (np. druków cyfrowych). Biurko Badacza dostępne będzie przez interfejs webowy.

Opracowane w Biurku Badacza materiały służyć mają do prowadzenia cyfrowych badań filologicznych w trybie indywidualnym lub zespołowym. W założeniu twórców BB powinno stać się podstawowym narzędziem wykorzystywanym w pracy m.in. nad grantami badawczymi.

Zadanie ukończone jest w 60% - aktualnie trwa łączenie w jedną całość modułów informatycznych.

W zadaniu zaangażowani są: **Prof. UAM dr hab. Bogdan Hojdis** (WFPiK) oraz **mgr Adam Cankudis** (WFPiK).

Zbiór narzędzi cyfrowych Węzła Filologicznego

System Digitalizacji i Analizy Korpusu (SDAK)

Usługa sieciowa dostępna przez interfejs webowy (docelowa integracja usługi w ramach Biurka Badacza) przeznaczona do automatycznego rozpoznawania jednostek metaforycznych w oparciu o Korpus Frazeologiczny Języka Polskiego opartego na archiwum prof. Stanisława Bąby.

Narzędzie wykorzystuje autorski **System Digitalizacji i Analizy Korpusu** (SDAK) opracowany siłami pracowników Wydziału Filologii Polskiej i Klasycznej UAM i Wydziału Matematyki i Informatyki UAM.

Zadanie ukończone w 60% - aktualnie trwa testowanie narzędzia i wprowadzanie poprawek zwiększających skuteczność rozpoznawania związków frazeologicznych.

W zadaniu zaangażowani są: **Prof. UAM dr hab. Jarosław Liberek** (WFPiK) oraz **Prof. UAM dr hab. Filip Graliński** (WMI).

Zbiór narzędzi cyfrowych Węzła Filologicznego

System Mapowania Tekstów Literackich (SMTL)

Usługa sieciowa dostępna przez interfejs webowy (docelowa integracja usługi w ramach Biurka Badacza) przeznaczona do ręcznego i automatycznego mapowania relacji zachodzących między postaciami, motywami, kontekstami w utworze literackim. W szerszej perspektywie narzędzie umożliwiać ma rozbudowywać poszczególne części mapy o nowe konteksty recepcyjne: religijne, literackie, psychologiczne, językoznawcze itp. oraz posiadać ma różnorodne modułowe narzędzia wizualizacji: grafy, schematy, listy, zakresy, typy źródeł oraz liczne odnośniki przydatne do analizy tekstu źródłowego (wszelkie dostępne źródła, dane kartograficzne, dane słownikowe itd.)

Zadanie ukończone w 80% - aktualnie trwa testowanie narzędzia i wprowadzanie poprawek zwiększających czytelność i funkcjonalność wyświetlanych struktur.

W zadaniu zaangażowani są: **Prof. UAM dr hab. Konrad Dominas** (WFPiK) oraz **mgr Adam Cankudis** (UAM).

Zbiór narzędzi cyfrowych Węzła Filologicznego

Internetowe Narzędzie do Kolaboratywnej Animacji i Hipertekstu (INKAH)

Usługa sieciowa umożliwiająca tworzenie, odczyt oraz wizualizację kolektywnie tworzonego hipertekstu z możliwością badawczej analizy aktywności badanej grupy.

Po pierwsze INKAH będzie rozbudowaną aplikacją sieciową do tworzenia utworów e-literackich. Po drugie, dzięki nakładce badawczej, umożliwi dostęp do aktywności użytkowników, dzięki której badacz będzie mógł obserwować rodzaje aktywności autorów i czytelników, sposoby współpracy zespołowej, sposoby i statystyki czytania.
Dzięki możliwości eksportu plików, INKAH stanie się też archiwum utworów w wolnym dostępie.

Zadanie ukończone w 60% - aktualnie trwa finalizacja etapu programowania i wprowadzanie poprawek zwiększających czytelność i funkcjonalność wyświetlanych hipertekstów.

W zadaniu zaangażowani są: **Prof. UAM dr hab. Elżbieta Winiecka** (WFPiK) oraz **dr Mariusz Pisarski** (Electronic Literature Organization).

Zastosowanie Węzła Filologicznego

- Wsparcie badaczy w ramach prowadzenia indywidualnych i kooperatywnych cyfrowych badań filologicznych (zespoły otwarte, zamknięte, lokalne, w ramach wielu ośrodków itd.).
- Przeniesienie części badań prowadzonych w sposób tradycyjny do intuicyjnej przestrzeni cyfrowej.
- Wsparcie badaczy w procesie realizacji i zarządzania grantami badawczymi.
- Umożliwienie tworzenia własnych, autorskich modułów informatycznych wspierających węższe zakresy badań i dołączanie tych modułów do Biurka Badacza.
- Cyfrowa publikacja wyników badań w formie publikacji, dokumentacji a także wsparcie w procesie udostępnienia narzędzi cyfrowych stworzonych w trakcie badań (baz, witryn, aplikacji itp.).
- Wsparcie promocji prowadzonych badań - indywidualnych i zespołowych.
- Zgodnie z założeniami projektu wszystkie narzędzia udostępniane będą bezpłatnie dla celów badawczych.

Dominas / Hojdis / Liberek / Pisarski / Winiecka / Wobalis

Podsumowanie

- Ramy czasowe projektu: 1.01.2021 - 31.12.2023
- Zasady udostępniania infrastruktury – w opracowaniu
- Infrastruktura UAM i infrastruktura Konsorcjum Dariah.lab
- Więcej: https://lab.dariah.pl/.
- O wszystkich modułach infrastruktury Dariah.lab w UAM można przeczytać np. tutaj: https://uniwersyteckie.pl/nauka/prof-katarzyna-klessa-nowa-era-humanistyki-z-dariahlab. Na stronach ŻU można też znaleźć dwa wywiady z Koordynatorami modułów SNJL, kolejne w przygotowaniu

uniwersyteckie.pl

Dziękujemy za uwagę

Prace finansowane ze środków projektu:
Cyfrowa infrastruktura badawcza dla humanistyki i nauk o sztuce DARIAH-PL
Projekt rozwojowy finansowany w ramach Programu Operacyjnego Inteligentny Rozwój na lata 2014-2020 (POIR.04.02.00-00-D006/20)

Fundusze Europejskie Inteligentny Rozwój

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego