Eszter Mihály – mihaly.eszter@oszk.hu

dHUpla infrastruktúra (Digital Humanities Platform – dhupla.hu)

A platform elsődleges célja, hogy a kutatók, olvasók számára egységes felületen, filológiai igényességgel hozzáférhetővé tegye a magyar kulturális örökség különböző intézményekben őrzött, eddig ismeretlen vagy méltatlanul elfeledett szöveges tartalmú, elsősorban kéziratos forrásait. Emellett olyan szerkesztőségi keretrendszert is kínál a tartalmak digitalizálásához, amely biztosítja a források egységes és színvonalas feldolgozását.

A rendszer rugalmas és moduláris, a tartalom előállítása többféle úton történhet, illetve több becsatlakozási pontot nyújt, amely az ábrán is végigkövethető. (A szaggatott nyíllal jelölt kapcsolatok a tervezett modul-fejlesztéseket jelölik.)

Az átírt szövegek a nemzetközileg támogatott TEI (Text Encoding Initiative) szabvány szerint annotált XML formátumban készülnek, amely a nemzetközi integráció mellett lehetővé teszi többek között a dokumentumok gépi feldolgozását és értelmezését (Linked Open Data), szemantikus hálók kiépítését (Semantic Web), adatgazdagítást (Data Enrichment), illetve a távoli olvasás (Distant Reading) különböző aspektusait. A szolgáltatás kapcsolatot létesít különböző névterekkel, bibliográfiai forrásadatbázisokkal, illetve nyelvi elemző szoftverekkel. Mindezek segítségével a legkülönbözőbb elemzések, korpuszlekérdezések, adatvizualizációk válnak megvalósíthatóvá.

Az infrastruktúra középpontjában a git verziókövető szoftver áll, amely végül teljes mértékben kiváltotta egy XML-adatbázis használatának szükségességét, nagyban leegyszerűsítve a rendszer használatát, karbantartását és fejlesztését. A dHUpla git-ben lévő források (szöveg, programkód) alapján publikál, így a git repository-k birtokában bárhol újraépíthető a teljes dHUpla honlap. A projektek mind önálló git repository-ban vannak, a publikáláshoz szükséges transzformációt docker container-ek végzik, minden egyes projekthez meg lehet adni saját ún. builder-t, amelyekben tetszőleges programnyelvet lehet használni. A HTML tartalmon túl Apache Solr index fájlt is előállítunk, amelynek segítségével a legkülönbözőbb facettált keresések is lehetővé válnak.

A publikációs felületen több módon lehetőség nyílik az átírt szöveg és az eredeti facsimile együttes vizsgálatára, a digitális objektumok különböző szempontú rendezésére, szűrések elvégzésére. Az egyes gyűjtemények megjelenésének, funkcióinak konfigurálása egyszerű szöveges fájlokban (yml) történik.

A kéziratok átírása során ezenkívül olyan kézírásfelismerő-modell épül, amely folyamatosan bővülve alkalmassá válik a magyar nyelvű kézírások automatikus felismertetésére, azaz mesterséges intelligencián alapuló gépi feldolgozására (Handwritten Text Recognition).

dHUpla user story

1. A felhasználó beszkenneli a publikálni kívánt dokumentumokat az előírt szabályok szerint (tiff, 600 dpi; a képeket a meghatározott konvenció szerint nevezi el és rendezi mappastruktúrába).
2. Létrejön a GitLabban egy gyűjteményt (collection). A gyűjtemények körét a felhasználó határozza meg. Egy gyűjtemény tartalmazhat egy vagy sok dokumentumot. A dokumentum egységét szintén a felhasználó határozza meg. (Lehet egy gyűjtemény egy regény, amely csak 1 dokumentumot tartalmaz (magát a regényt), de lehet gyűjtemény egy levelezés is, amely több száz dokumentumot tartalmaz, ha 1 levél 1 dokumentum.)
A gyűjtemény létrehozója a tulajdonos, de megoszthatja gyűjteményét más regisztrált felhasználóval, különböző jogosultságokat kiosztva (főszerkesztő, szerkesztő, néző).
3. A gyűjtemény létrehozásakor a felhasználó opcionálisan létrehoz egy Kanban board-ot a projekt (gyűjtemény) nevével és a munkafolyamatnak megfelelő oszlopokkal.
4. A felhasználó feltölti a szkennelt képeket az erre kijelölt storage-ra, az előre meghatározott mappastruktúra szerint (1 dokumentum mindig 1 mappa). A feltöltés után a master képek megfelelő helyre történő archiválása, illetve biztonságos tárolásának megoldása automatikusan történik. Ezzel egyidőben a rendszer minden képből generál egy ún. munkafájlt is, kisebb felbontásban (png/jpg, 300 dpi), ugyanolyan mappastruktúrában, mint a master fájl esetében, ezeket a felhasználó betölti a megfelelő git repositoryba (project-assets).
5. Amennyiben a felhasználó Transkribus klienst használ a szövegátíráshoz, illetve a zónázáshoz, feltölti oda a munkafájlokat. A Transkribus választása esetén a rendszer biztosítja a Transkribus szerveréről való folyamatos biztonsági mentést (Page XML). Ha a felhasználó más szövegátíró, illetve zónázó eszközt használ, akkor ezt a lépést kihagyja. A felhasználó dönthet úgy is, hogy kezdetektől a dHUpla Oxygen frameworkjében szerkeszti a TEI XML fájlokat, ekkor a zónázáshoz az Oxygen framework Image Map editor eszközét használhatja (vagy kihagyja a text-image linking fázisát, és a facsimiléket csupán oldalanként behivatkozza).
6. Amennyiben a felhasználó eddig nem dHUpla Oxygen frameworköt használt, a Transkribusból, illetve egyéb átíró és zónázó eszközökből exportálja/menti a dokumentumot TEI XML formátumban. A TEI XML git-be való feltöltésekor a felhasználó hibaüzenetet kap (de a rendszer engedi a feltöltést), ha az aktuális TEI séma szerint nem valid TEI XML fájlokat próbál feltölteni.
7. Ezután a dHUpla Oxygen frameworkjében történhet a TEI XML további szerkesztése, javítása, metaadatolása, adatgazdagítása, filológiai feldolgozása, amely megkönnyíti a dHUpla-szabványhoz való igazodást is. A felhasználó dönthet úgy is, hogy nincs szüksége további szerkesztésre, illetve más XML szerkesztőben végzi a további feladatokat. Ha a dokumentumo(ka)t máshol szerkeszti tovább, akkor az újabb TEI XML-(eke)t mindig újból feltölti.
8. A szerkesztés során a felhasználó folyamatosan megtekintheti a dokumentumo(ka)t egy tesztkörnyezetben, ami az éles környezet pontos mása. Ha a dokumentumot/szövegkorpuszt publikálhatónak véli, felterjeszti a dHUplán való publikálásra. A publikálási folyamatot a jogosultsági körök szabályozzák. A felterjesztés pozitív elbírálása után történik az éles környezetben való publikáció, egyszerűen a dokumentum státuszának átállításával. A felhasználónak lehetősége van a dokumentumok/gyűjtemények nyilvánosságból való visszahívására is.

Eszter Mihály – mihaly.eszter@oszk.hu

dHUpla infrastructure (Digital Humanities Platform – dhupla.hu)

The goal of the platform is to make accessible to researchers and readers the hitherto unknown or undeservedly forgotten textual sources of the Hungarian cultural heritage mainly manuscripts, preserved in various institutions, on a uniform interface and with philological sophistication. In addition, it also offers an editorial framework for the digitization of content, which ensures a uniform and high-quality processing of sources.

The system is flexible and modular, the content can be produced in several ways, and it provides several connection points, which can also be followed on the poster. (The links marked with dashed arrows indicate planned module developments.)

The transcribed texts are prepared in an annotated XML format according to the internationally supported TEI (Text Encoding Initiative) standard, which, in addition to international integration, enables the machine processing and interpretation of documents (Linked Open Data), the construction of semantic networks (Semantic Web), data enrichment and various aspects of distant reading. The service establishes connections with various entity databases, bibliographic source databases, and language analysis software. All of this makes a wide variety of analyses, corpus queries, and data visualizations become feasible.

At the heart of the infrastructure is the git version tracking software, which eventually completely eliminated the need to use an XML database, greatly simplifying the use, maintenance and development of the system. dHUpla publishes from sources (text, program code) in git, so the entire dHUpla website can be rebuilt anywhere with the git repositories. The projects are all in a separate git repository, the transformation required for publishing is done by docker containers, and each project can be assigned its own builder, in which any programming language can be used. In addition to the HTML content, we also produce an Apache Solr index file, which enables a wide variety of faceted searches.

On the publication interface, it is possible to examine the transcribed text and the original facsimile together in several ways, to sort the digital objects according to different aspects, and to carry out filters. The appearance and functions of each collection are configured in simple text files (yml).

In the process of transcribing the manuscripts, a handwriting recognition model is also built, which continuously expands and becomes suitable for automatic recognition of Hungarian handwriting, i.e. machine processing based on artificial intelligence (Handwritten Text Recognition).

ORSZÁGOS SZÉCHÉNYI KÖNYVTÁR

dHUpla user story

1. The user scans the documents to be published according to the prescribed rules (tiff, 600 dpi; the images are named according to the defined convention and arranged in a folder structure).

2. A collection is being created in GitLab. The scope of the collections is defined by the user. A collection can contain one or many documents. The document unit is also defined by the user. (A collection can be a novel that contains only 1 document (the novel itself), but it can also be a collection of correspondence that contains hundreds of documents, if 1 letter is 1 document.)

The creator of the collection is the owner, but they can share the collection with other registered users, assigning different rights (chief editor, editor, viewer).

3. When creating the collection, it is optional for the user to create a Kanban board with the name of the project (collection) and the columns corresponding to the workflow.

4. The user uploads the scanned images to the designated storage, according to the predefined folder structure (1 document is always 1 folder). After uploading, the master images are automatically archived in the appropriate location and securely stored. At the same time, the system also generates a working file, in a lower resolution (300 dpi, png/jpg), in the same folder structure as in the case of the master file, the user loads these into the appropriate git repository (project-assets).

5. If the user works with a Transkribus client for text transcription and zoning, they upload the work files there. If they choose Transkribus, the dHUpla-system provides a continuous backup from the Transkribus server (Page XML). In the case of using another text transcription or zoning tool, this step is skipped. The user can also decide to edit the TEI XML files in dHUpla's Oxygen framework from the beginning, in which case they can use the Oxygen framework's Image Map editor tool for zoning (or skips the text-image linking phase and references images per page).

6. If the user has not used dHUpla's Oxygen framework so far, they export/save the document in TEI XML format from Transkribus or other transcribing and zoning tools. When uploading TEI XML to git, the user receives an error message if they try to upload TEI XML files that are not valid according to the current TEI schema (but the system allows the upload).

7. The TEI XML can then be further edited, corrected, metadata added, data enriched, and philologically processed in dHUpla's Oxygen framework, which also facilitates alignment with the dHUpla standard. The user can also decide that they do not need further editing, or that they perform the additional tasks in another XML editor. If they continue to edit the document(s) elsewhere, the newer TEI XML(s) will always be uploaded again.

8. During editing, the user can continuously view the document(s) in a test environment which is an exact copy of the live environment. If they consider the document/text corpus to be publishable, they submit it for publication. The publishing process is regulated by authorization circles. After a positive evaluation of the submission, the publication takes place in the live environment simply by changing the status of the document. The user also has the option of recalling documents/collections from the public.