Irská národní folklorní sbírka:
jak (ne)zdigitalizovat 300 000 rukopisných stránek

Michal Měchura, UČO 462258, michmech@mail.muni.cz

Irská národní folklorní sbírka (irsky *Cnuasach Bhéaloideas Éireann,* anglicky *National Folklore Collection,* dále INFS) sídlí v University College Dublin (UCD) a je to jedna z největších sbírek folklorního materiálu na světě: obsahuje 2 miliony rukopisných stran, 12 000 hodin zvukových nahrávek a 80 000 fotografií. Dokumentováním lidového umění, lidových tradic a ústní historie v Irsku i v okolních gaelských zemích (Skotsko a ostrov Man) se zabývá už od svého založení v roce 1935. Navzdory své velikosti, věku i prestiži v očích irské veřejnosti – nebo možná právě proto – vstoupila do digitálního věku relativně pozdě: až donedávna byla pouze fyzickou institucí, veřejnosti nabízela možnost nahlédnout do svých sbírek jen osobní návštěvou a většina metadat se udržovala v papírových kartotékách. To se změnilo v roce 2012, kdy se ve spolupráci s Dublin City University (DCU), která už měla zkušenosti s digitalizací v humanitních vědách, pustila do rozsáhlého digitalizačního projektu financovaného grantem od vlády Irské republiky.[1] Jeho výstupem se staly internetové stránky `www.duchas.ie`,[2] které od roku 2013 umožňují veřejnosti nahlédnout do stále rostoucí zdigitalizované podmnožiny materiálů držených v INFS.

Úkolem mé eseje je seznámit čtenáře s tímto digitalizačním projektem, kterého jsem se sám zúčastnil v roli technického architekta a jsem tedy zodpovědný, v dobrém i zlém smyslu, za jeho informačně-technologické provedení.[3] Nejdříve podrobněji vysvětlím, co se vlastně v projektu digitalizovalo a jak. Potom se (sebe)kriticky zastavím nad některými aspekty, které si podle mého soudu zaslouží pozornost, a zhodnotím, jak uspokojivě nebo neuspokojivě byly v projektu vyřešeny.

Školní sbírka

Digitalizační projekt INFS, který trvá dodnes, se ve své dnes už završené první fázi zaměřil na jednu konkrétní podmnožinu INFS, a sice na tzv. Školní sbírku (irsky *Bailiúchán na Scol,* anglicky *The Schools' Collection*). Školní sbírka sestává z rukopisných materiálů a neobsahuje žádná jiná média. V současné době (podzim 2016) se rozbíhá nová fáze projektu, ve které se budou

1 Konkrétně od Ministerstva pro umění, kulturní dědictví, regiony, venkov a irsky mluvící území.

2 Podstatné jméno *dúchas* v irštině znamená *původ, (kulturní) dědictví.*

3 Dodnes s projektem spolupracuju jako externí konzultant.

digitalizovat další podmnožiny INFS včetně fotografií a zvukových nahrávek, těmi se zde však zabývat nebudu. Zde se pro jednoduchost soustředím jen na digitalizaci Školní sbírky.

Školní sbírka vzešla z projektu, který se uskutečnil v letech 1937 až 1938 na všech základních školách Irské republiky (to znamená s vyloučením Severního Irska). Žáci škol dostali za úkol nasbírat od dospělých příbuzných folklorní materiál na různá témata: místní legendy, popisy lidových oděvů, řemesel a společenských událostí, domněnky o původu místopisných názvů a podobně. Pod vedením svých učitelů je žáci sepsali do sešitů, které se potom poslaly do INFS v Dublinu a kde jsou dodnes k nahlédnutí. Byl to jeden z nejrozsáhlejších projektů na sběr folkloru na světě a jeho výsledkem je jakási ‚fotografie' irského lidového života dětskýma očima před Druhou světovou válkou, tedy v době, kdy většina irského obyvatelstva ještě žila na venkově a kdy irský veřejný život nebyl tak výrazně poangličtěn jako dnes.

Fyzicky se Školní sbírka skládá ze 4 413 sešitů svázaných do 1 124 svazků. Celkový počet stránek je 362 077. Každý jednotlivý sešit pochází z jedné konkrétní školy, jejíž název a sídlo je uvedeno na titulní stránce. Na obsahu každého sešitu se typicky podílelo několik žáků najednou, z nichž každý napsal jeden nebo více článků. U každého článku bývá uvedeno jméno a další biografické údaje sběratele (= žáka, který článek napsal) a informátora nebo informátorů (= dospělých osob, od kterých sběratel informace vyslechl a zapsal). Text je bez výjimky rukopisný, místy protkaný ilustracemi. Zhruba dvacet procent obsahu je v irštině, zbytek v angličtině.

Jak se digitalizovala Školní sbírka

Školní sbírku lze chápat jako knihovnu takzvané *šedé literatury:* je to velké množství materiálu, který nikdy nevyšel ani nevyjde formou publikace a který slouží spíš jako surový materiál ke studiu. Smyslem digitalizačního projektu bylo udělat z této analogické knihovny knihovnu digitální.

Všechny stránky všech sešitů byly automatizovaně oskenovány a uloženy ve čtyřech verzích: jedna verze s vysokým rozlišením pro dlouhodobou archivaci a další tři, v různých velikostech a rozlišeních, pro zobrazování na internetu. Na ambici digitalizovat text jsme rezignovali velmi brzo, a to ze zřejmých důvodů: OCR na rukopis asi 50 000 různých autorů prostě použít nelze a manuální opis tak velkého počtu stránek by byl prohibitivně drahý.[4] Soustředili jsme se tedy na pouhou indexaci metadat. I ta byla náročná: tým anotátorů (jehož velikost a složení se v průběhu projektu měnily) tuto práci začal v roce 2012 a dokončil ji teprve letos (2016). Práce postupovala

4 Tento problém jsme později obešli tím, že jsme přizvali dobrovolníky z řad veřejnosti. Zdigitalizovaná INFS od roku 2015 používá crowdsourcing a dobrovolníci už ručně opsali přes 40 000 stránek (stav z podzimu 2016).

geograficky, od jednoho hrabství k druhému, a výsledky se průběžně zveřejňovaly na stránkách `www.duchas.ie`.

Metadata zaznamenáváme na dvou úrovních: jednak k sešitům, jednak k jednotlivým článkům uvnitř sešitů. K sešitům jsou metadata poměrně úsporná: sestávají z názvu a geografické polohy školy, ze které sešit pochází a dále ze jména a dalších biografických údajů o učiteli, který sepisování sešitu koordinoval. K článkům jsou metadata mnohem bohatější. Obsahují:

- Nadpis. Některé články ale nemají nadpis žádný nebo je příliš generický; v takovém případě se zaznamenává i úryvek z textu, obvykle první věta nebo dvě.
- Téma článku, například řemesla, oděv, legendy, lidové slavnosti…
- Místo, kterého se článek týká, pokud se liší od umístění školy.
- Jména a další biografické údaje o sběrateli a informátorech: věk, profese, geografická poloha. U informátorů se zaznamenává i jejich vztah ke sběrateli: rodič, prarodič, jiný příbuzný, cizí osoba.
- Údaj o tom, v kterém jazyku článek je: irsky, anglicky, smíšeně.
- Údaj o tom, na kterých stránkách kterého sešitu se článek nachází. Stránky a články si navzájem neodpovídají jedna ku jedné: stránka může obsahovat víc než jeden článek a článek se může rozkládat na víc než jedné stránce. Rozlišujeme tedy mezi fyzickou (stránky) a logickou (články) strukturou každého sešitu.

Úkolem členů indexačního týmu je každý článek zběžně pročíst a zaznamenat k němu metadata. Indexování metadat probíhá v nástroji jménem *Léacslann* (Měchura 2012). Léacslann je obecný nástroj na práci se strukturovanými záznamy dat, například se slovníkovými hesly. Vznikl v jedné ze zúčastněných univerzit (DCU) pro potřeby kompilace různých databází v oboru lexikografie, terminologie a humanitních věd. Pro potřeby konkrétního projektu lze Léacslann přizpůsobit vytvořením takzvané *aplikace* uvnitř něj. Indexování Školní sbírky je jedna z takových aplikací; typická obrazovka této aplikace je k nahlédnutí ve **vyobrazení 1**.

Veřejnosti se výstup digitalizačního projektu prezentuje na stránkách `www.duchas.ie`, viz **vyobrazení 2**. Stejně jako Léacslann (respektive aplikace v něm) byly tyto stránky vytvořeny na míru a takzvaně ‚na zelené louce', bez použití existujícího repozitáře pro digitální knihovny (jako Fedora[5] nebo DSpace[6]) ani žádné jiné generického platformy. Léacslann i veřejné stránky jsou

5 `http://www.fedorarepository.org/`

6 `http://www.dspace.org/`

webové aplikace vytvořené v *Microsoft .NET Framework* (konkrétně v jazyce C#), data se ukládají jednak v databázi *Microsoft SQL Server* (metadata), jednak obyčejně na disku (oskenované stránky).

Dlouhodobou archivaci oskenovaných stránek (verzí s vysokým rozlišením) a získaných metadat má na starosti jedna ze zúčastněných univerzit (UCD), která se zavázala zajistit ji i v případě, že v budoucnosti nebude pokračovat financování projektu. Veřejně přístupný web `www.duchas.ie` však takto zajištěn není, jeho dlouhodobé hostování a udržování v chodu je závislé na nepřetržitém financování.

Standardy a interoperabilita

Ve Školní sbírce se metadata každého jednotlivého článku ukládají jako jeden (relativně malý) XML dokument. Struktura těchto XML dokumentů byla navržena speciálně pro potřeby tohoto projektu a zcela bez ohledu na existující metadatové standardy, což lze – ne neprávem – považovat za hlavní slabinu projektu. Ukázka takového dokumentu je ve **vyobrazení 3**. Čtenář si jistě všimne, že jména XML elementů a jejich vnořovací hierarchie neodpovídají žádnému známému standardu. Třem vybraným aspektům struktury těchto metadat se nyní budeme věnovat trochu detailněji.

Jména lidí

Již v samém začátku projektu bylo rozhodnuto, že jména lidí (= sběratelů, informátorů, učitelů) se budou v metadatech zmiňovat doslova, nikoli odkazem na číselník nebo na externí soubor autorit. Pokud je tatáž osoba zmíněna na více místech, například jako sběratel více než jednoho článku, údaje se v obou článcích duplikují. To je nutné zlo, které jsme se rozhodli přijmout, protože rezoluce totožnosti osob zmiňovaných ve Školní sbírce by byla nad naše síly, hlavně kvůli jejich vysokému počtu: kolem 50 000 sběratelů a zhruba stejný počet informátorů.

Další nepohodlný kompromis je ve způsobu, jak se v metadatech Školní sbírky jména kódují. Irská osobní jména v irštině i v angličtině se vyznačují variabilitou, která nemá jinde v Evropě obdobu (viz např. Nic Cóill 2011). Jedno jméno lze uvést mnoha různými způsoby, většinu domorodých jmen lze i překládat z jazyka do jazyka a tatáž osoba se může v různých kontextech vyskytovat pod různými variantami téhož jména. Tato variabilita byla ve třicátých letech dvacátého století ještě nespoutanější než dnes. V ideálním případě bychom tedy museli každé jméno podrobně označkovat a převést na jeho základní tvar, jak to předepisuje například standard TEI. I to jsme však uznali za úkol mimo naše možnosti a místo toho jsme se rozhodli každé jméno v metadatech

citovat doslova tak, jak je zmíněno v rukopisném textu. Jediné značkování, které se na jménech provádí, je vyznačení místa, kde začíná část relevantní pro abecední řazení (což u irských jmen není vždy totožné s celým příjmením: příjmení začínající na *Ó* nebo *Mac* se třídí abecedně až podle toho, co následuje).

Číselník témat

U většiny článků je vyznačeno, jakého tématu se týkají: řemesel, oděvů, legend, lidových slavností a podobně. Na rozdíl od jmen lidí se tato informace neukládá doslova, nýbrž odkazem na číselník témat, který obsahuje 221 témat zorganizovaných do hierarchického stromu s devíti vrcholy. Tento strom témat byl sestaven pro potřeby Školní sbírky a vychází z původního zadání, jak ho autoři navrhli ve třicátých letech dvacátého století. Neodpovídá tedy žádnému externě existujícímu schématu na klasifikaci materiálu podle témat. Otázka je, zda by se měl na některý externí standard převést nebo namapovat, a pokud ano, na který.

Jeden potenciální kandidát je kniha *A Handbook of Irish Folklore* (Ó Súilleabháin 1970), která obsahuje katalog témat sepsaný speciálně pro studium irského folkloru. V Irsku platí za folkloristickou ‚Bibli'. Vznikla však až několik desetiletí po Školní sbírce a je na ní nezávislá. Navíc není k dispozici ve strojově srozumitelné podobě.

Další kandidát je schéma zvané *Aarne-Thompson* (Aarne a Thompson 1987), podle kterého se často katalogizuje při studiu evropských folklorních tradic. Je to hierarchický katalog typů lidových zkazek, který je navržen tak, aby pokryl všechny folklorní tradice, které se v Evropě běžně vyskytují. Existuje i jeho mapování na irské poměry (Ó Súilleabháin a Christiansen 1967). Tento katalog témat je k dispozici ve strojově srozumitelné podobě a v současné době se zvažuje jeho vpracování do zdigitalizované Školní sbírky.

Georeference

Jediným aspektem digitalizované Školní sbírky, o kterém se dá říct, že je dobře interoperatibiltí s externím standardem, je aspekt georeference. Georeferencí se myslí údaj o geografii škol (kde se nacházejí), lidí (kde bydlí, odkud původně jsou) a článků (jakého místa se týkají).

Tyto údaje se ve Školní sbírce zaznamenávají odkazem na Databázi místopisných názvů Irska (irsky *Bunachar Logainmneacha na hÉireann,* anglicky *Placenames Database of Ireland,* dále DMNI), která v Irsku platí za standard. DMNI je projekt, který se už mnoho let uskutečňuje na Dublin City University (DCU) jako zakázka pro vládu Irské republiky. Veřejným výstupem jsou internetové

stránky `www.logainm.ie`, kde jsou veškeré údaje k dispozici i ve strojově srozumitelné podobě, a to jak ve vlastním interním XML formátu, tak jako množiny tripletů RDF. Každá geografická jednotka v Irsku, která má jméno, má v DMNI své vlastní neměnné URL (a v RDF má URI). Mnoho geografických jednotek je odsud dále namapováno (relací OWL *same as*) na externí databáze jako Geonames a WikiData. Na tato neměnná URL se odkazují georeferenční údaje ve Školní sbírce.

Závěr

Internetové stránky `www.duchas.ie` jsou digitální knihovna šedé literatury, která uspokojuje hlad irské veřejnosti po folklorním materiálu, který byl do té doby přístupný jen osobní návštěvou v INFS v Dublinu. Je to hojně navštěvovaný server, každý měsíc obslouží kolem půl milionu shlédnutí. V tom směru tedy projekt svůj úkol víc než splnil. Smyslem digitalizačních projektů ale není jen uspokojit okamžité potřeby veřejnosti, je jím i zajistit počítačovou zpracovatelnost zdigitalizovaného materiálu a jeho dlouhodobé uchování. Absence externě uznávaných standardů v metadatech tento smysl částečně potírá. Tato slabina je však správcům zdigitalizované Školní sbírky známa a pravděpodobně bude v budoucnu odstraněna převedením nebo alespoň namapováním existujících metadat na standardy.

Literatura

Aarne, Antti; Thompson, Stith (1987) *The Types of the Folktale: A Classification and Bibliography.* Second Revision. Helsinki: Suomalainen Tiedeakatemia.

Nic Cóil, Róisín (2011) 'Irish prefixes and the alphabetization of personal names'. *The Indexer* 29(2): C1-C6.

Měchura, Michal (2012) 'Léacslann: a platform for building dictionary writing systems'. *Proceedings of the 15th Euralex International Congress:* 855-861.

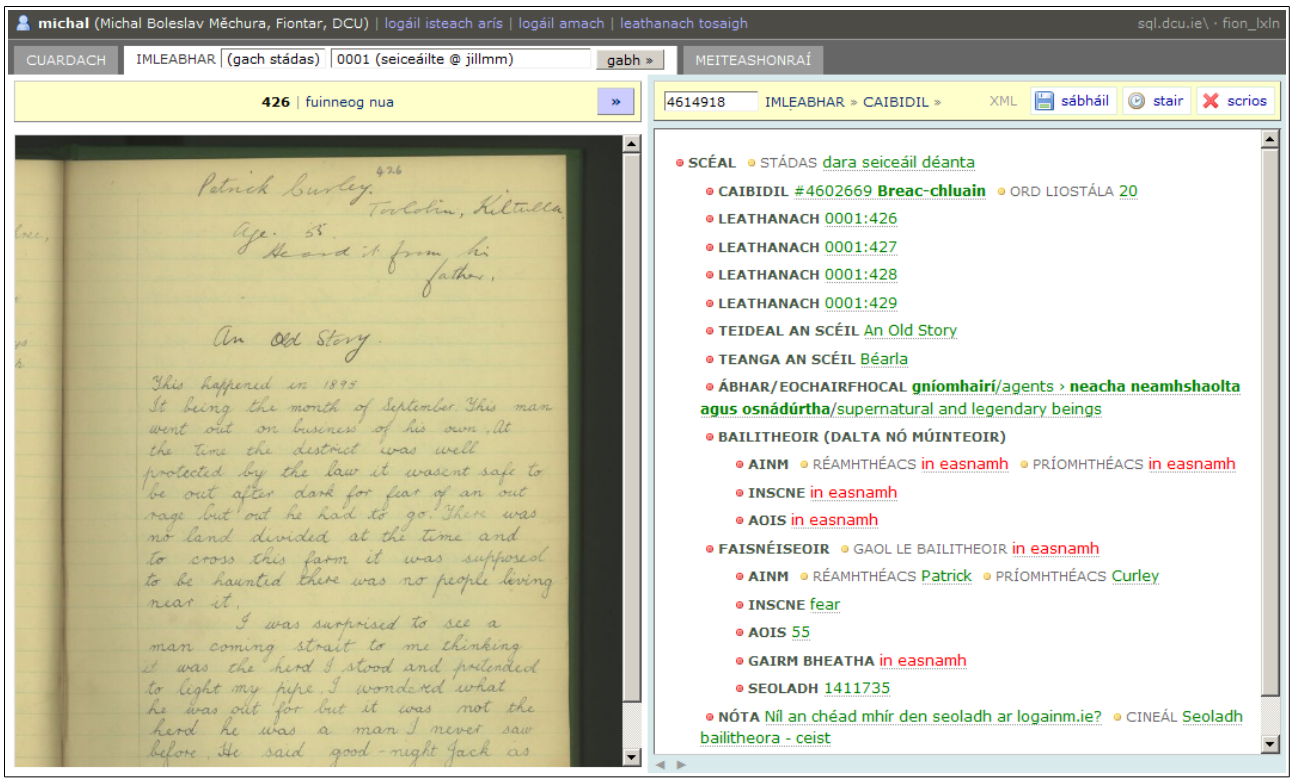
Ó Catháin, Séamas (1988) 'Súil siar ar Scéim na Scol 1937-1938' [retrospektivní pohled na Školní sbírku 1937-1938]. *Sinsear* 5: 19-30.

Ó Súilleabháin, Seán; Christiansen, Rheidar Th. (1967) *The Types of the Irish Folktale.* Helsinki: Suomalainen Tiedeakatemia.

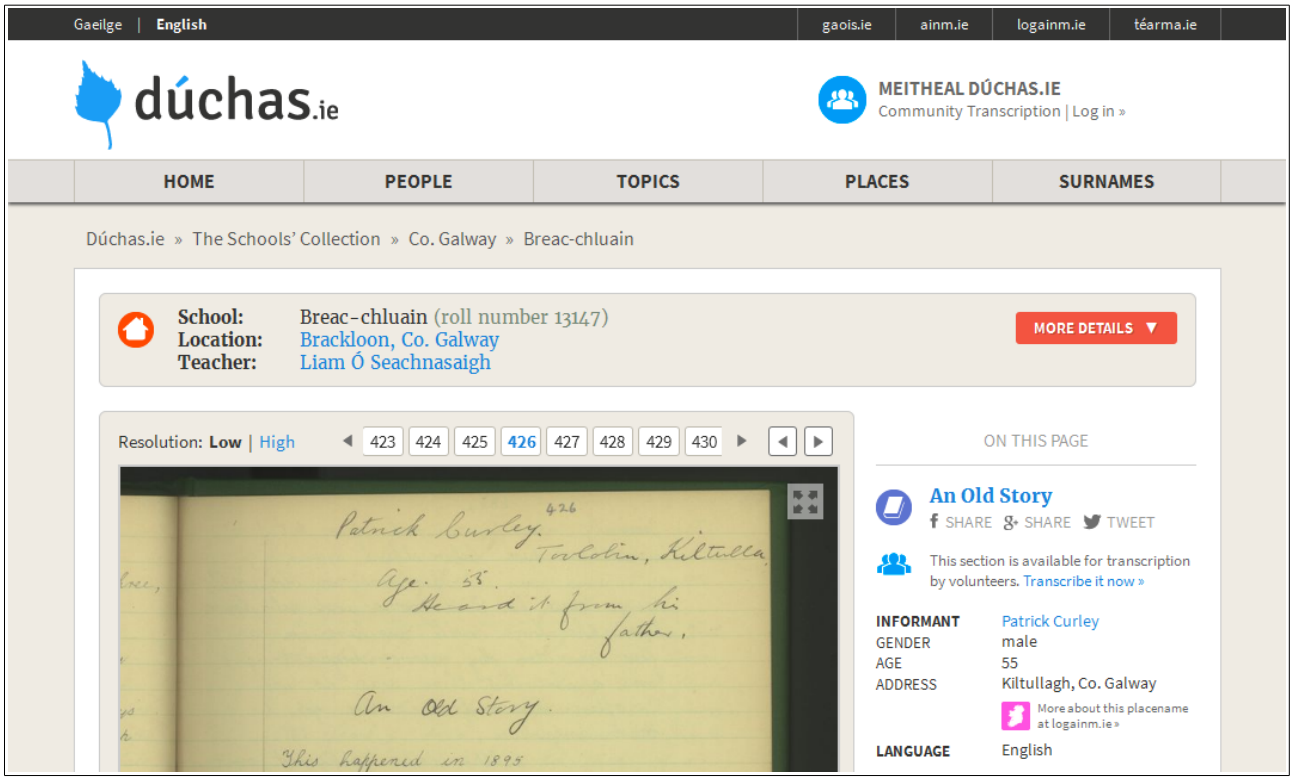
Ó Súilleabháin, Seán (1970) *A Handbook of Irish Folklore.* Detroit: Singing Tree Press.

Metadata o této eseji *(v nekvalifikovaném Dublin Core)*

creator	Michal Měchura
title	Irská národní folklorní sbírka: jak (ne)zdigitalizovat 300 000 rukopisných stránek
description	Esej o digitalizaci Irské národní folklorní sbírky.
date	2016-12-05
language	cs
subject	folklor
subject	Irsko
subject	digitalizace
subject	digitální knihovny
subject	digitální humanitní vědy



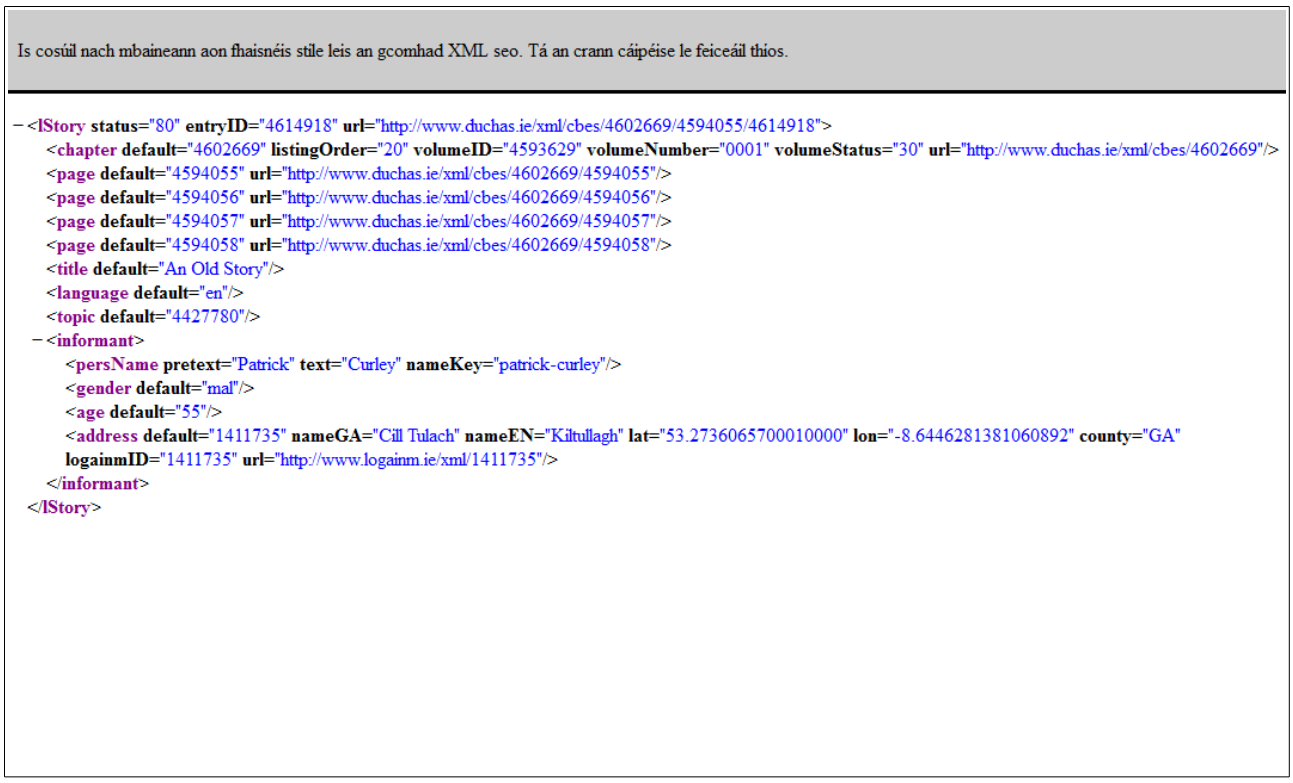
Vyobrazení 1: Aplikace pro indexaci Školní sbírky v systému Léacslann.



Vyobrazení 2: Stránka a metadata článku na ní tak, jak se zobrazují veřejnosti.

URL této stránky irsky: `http://www.duchas.ie/ga/cbes/4602669/4594055/4614918`

URL této stránky anglicky: `http://www.duchas.ie/en/cbes/4602669/4594055/4614918`



Vyobrazení 3: Strojové čitelná a veřejně přístupná (pod licencí CC-BY-NC) metadata jednoho článku. URL této stránky: `http://www.duchas.ie/xml/cbes/4602669/4594055/4614918`