STANDARDY PRO ZACHYCENÍ VÝSLEDKŮ ROZPOZNÁNÍ TEXTU

Boris Lehečka; lehecka@mzk.cz; (Moravská zemská knihovna v Brně)

Článek se zaměřuje na standardy pro ukládání výsledků rozpoznání tištěného (OCR), popř. rukopisného (HTR) textu. Srovnává několik aktuálně používaných formátů na bázi značkovacího jazyka XML i JSON. Hlavní pozornost se soustředí na zachycení údajů o rozpoznaném jazyce, toku textu na stránce a mezi stránkami, včetně problematiky dělení slov na konci řádku.

http://doi.org/10.52036/1335793X.2023.SC2.63-74

ÚVOD

Optické rozpoznání textu neboli OCR (Optical Character Recognition) a nověji také rozpoznání rukopisu neboli HTR (Handwritten Text Recognition) představují významnou součást zpřístupňování písemného kulturního dědictví. Díky těmto technologiím mohou písemnosti existovat v digitální podobě a vstupovat do interakce na nejrůznějších úrovních: od čtenářů (např. pro čtení v čtečkách elektronických knih) až po badatele (příprava knižní edice, budování korpusů, stylometrická analýza ap.).

Rozpoznání textu na stránce zahrnuje nejen identifikaci samotného textu, ale i dalších objektů nebo jednotlivých částí, které se na stránce, tj. fyzickém nosiči, vyskytují, jako jsou obrázky, tabulky, živá záhlaví a zápatí apod. Aby bylo možné tyto informace zahrnout do různých fází zpracování rozpoznávaného dokumentu (např. pro trénování špatně rozpoznaných znaků nebo export do různých formátů), vznikaly pro tyto potřeby v rámci jednotlivých softwarů interní datové formáty a následně jejich strojově čitelná reprezentace určená ke sdílení, případně dalšímu zpracování.

Formáty určené ke sdílení zachovávají vazbu mezi rozpoznanými prvky (obrázky, odstavci, slovy apod.) a vstupním obrázkem. To lze využít při další práci s digitální podobou: např. nalezený výraz v rozpoznaném textu je možné přesně ohraničit v digitální kopii konkrétní stránky. Míra detailu a způsob zachycení jednotlivých částí digitalizovaného dokumentu se postupně proměňovaly na základě potřeb jednotlivých subjektů a konsorcií, popř. na základě rozvoje nových technologií. Kromě samotné přítomnosti textu lze podchytit také jeho kvality, jako je písmo a jeho řez, nebo rozdělení slova na konci řádku pomocí spojovníku. Pro tyto účely vzniklo několik hlavních formátů, které využívají obecný značkovací jazyk XML, v poslední době, zejména v souvislosti s webovými službami pro rozpoznání textu, se objevily standardy využívající formát JSON.

Následující analýza se zaměří na existující formáty pro sdílení výsledků rozpoznání textu, zejména pokud jsou využívané v digitálních knihovnách založených na systému Kramerius (Kramerius, bez data), viz https://registr.digitalniknihovna.cz. Budeme se věnovat převážně těm jejich součástem, které mají dopad na zachycení textu a jeho kvalit. Při analýzách jsme využívali dostupné ukázky v repozitářích na GitHubu a dokumenty z digitálních knihoven Knihovny Akademie věd ČR a Národní knihovny ČR. Článek doprovází repozitář na GitHubu s ukázkami a doplňujícími informacemi (Lehečka, 2023).

ALTO (ANALYZED LAYOUT AND TEXT OBJECT)

Standard pro analýzu layoutu a textu, který je znám pod zkratkou ALTO (Library of Congress, 2023), začal vznikat v roce 2000, nyní jeho vývoj spravuje Knihovna Kongresu.

Standard používá sémantické verzování, tj. při podstatné změně, která vede ke ztrátě zpětné kompatibility s verzí předchozí, se mění hlavní (major) číslo verze, spolu s tím dochází ke změně jmenného prostoru všech elementů XML. Při menších změnách se mění pouze označení variantní (minor) verze. Od verze 4.0 je specifikace ALTO dostupná pod licencí CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/). Poslední verze 4.4 pochází z března 2023.

Webová stránka Knihovny Kongresu obsahuje přehled všech verzí s bodovým popisem hlavních rozdílů a odkazem na definici schématu ve formátu XSD. Dále obsahuje základní popis hlavních elementů, které však vycházejí z verze 2.0 a nereflektují novější změny, např. volitelnost atributu *@FONTSIZE*. Součástí stránek jsou i odkazy na další zdroje, jako je složení ediční rady, zápisy z jejího setkání (pouze do roku 2015), zdrojové kódy, dokumentace a další dokumenty ke standardu v repozitáři GitHubu (https://github.com/altoxml) nebo seznam institucí (knihoven), které tento standard implementovaly, a přehled softwarových aplikací a nástrojů, které s tímto formátem dokážou pracovat (generovat jej, zobrazovat, transformovat).

I když je tento standard doporučován jako rozšiřující schéma v rámci standardu METS (Metadata Encoding and Transmission Schema) (Library of Congress, 2023), a to v sekci pro administrativní metadata, používá se často i samostatně jako výstup z programu pro rozpoznání textu.

Dokument ALTO XML zachycující digitální předlohu je rozdělen na několik základních částí: nastavení (měrné jednotky, metadata); styly používané strukturními prvky na stránce (např. písmo, zarovnání odstavce); značky definující vlastnosti, na něž je možné odkázat (např. pojmenované entity, role); pořadí čtení segmentů na stránce, pokud není kontinuální; samotná stránka tvořená okraji a potištěnou plochou, na níž se vyskytují jednotlivé rozpoznané části (bloky textu, ilustrace, grafické elementy apod.).

Definice standardu je zajímavá i z formálního hlediska, a to nejednotností v pojmenování elementů a atributů. Kořenový prvek používá malá písmena **(<alto>**), další elementy začínají převážně velkým, ale i malým písmenem a používají velbloudí notaci, tzv. camelCase, např. **<Layout>**, **<MeasurementUnit>**, **<sourceImageInformation>**, **<fileName>**; objevují se ale i elementy tvořené majuskulemi: **<ALTERNATIVE>**, **<SP>** a **<HYP>**. Názvy atributů jsou tvořeny velkými písmeny (*@CONTENT*, *@HEIGHT*), pro oddělení víceslovných komponent se používá podtržítko (*@PRINTED_IMG_NR*). Také zde se však najdou výjimky, např. *@fileIdentifierLocation* nebo *@SCHEMAVERSION*. Nejednotnost se projevuje např. i v definicích typu některých atributů. Poslední verze standardu, 4.4, obsahuje 58 elementů a 58 atributů.

Měrné jednotky je potřeba definovat jako 1/10 mm nebo 1/1200 palce. Aby bylo možné přepočítávat hodnoty uváděné u jednotlivých elementů (např. umístění bloku textu) v případě, že se pro prezentaci použije obraz s jiným rozlišením, je možné koordináty převést na pixely podle definovaných vzorců.

Mezi hlavní prvky XML pro zachycení rozpoznaného textu a dalších objektů patří následující:
Element **<Description>** obsahuje metadata o zdrojovém souboru (**<sourceImageInformation>**) a způsobu jeho zpracování (**<OCRProcessing>** a nověji obecnější **<Processing>**) a o měrných jednotkách použitých ve výsledném dokumentu ALTO (**<MeasurementUnit>**).

Element **<Styles>** obsahuje prvky pro definici stylů používaných na stránce, a to zvlášť na úrovni odstavců (**<ParagraphStyle>**) a na úrovni úseků textu (**<TextStyle>**), na tyto styly se z objektů na stránce odkazuje pomocí identifikátoru. Podřazené objekty v tomto případě dědí vlastnosti z objektů nadřízených. Tento element je nepovinný, takže nástroje OCR nemusejí typografické kvality textu zachycovat, což je mj. případ českého projektu PERO OCR (PERO OCR, bez data).

Na úrovni formátování odstavců lze zachytit následující vlastnosti: zarovnání (*@ALIGN*, tj. vlevo, vpravo, na střed, do bloku), odsazení zleva (*@LEFT*), popř. zprava (*@RIGHT*), řádkování (*@LINESPACE*) a odsazení prvního řádku (*@FIRSTLINE*).

Pro podchycení formátování textových úseků jsou k dispozici tyto vlastnosti: název fontu (*@FONTFAMILY*, např. Times New Roman), typ (*@FONTTYPE*, tj. patkový, bezpatkový), šířka (*@FONTWIDTH*, tj. pevná nebo proporcionální), velikost (*@FONTSIZE*; v bodech, tj. 1/72 palce), barva (*@FONTCOLOR*) a styl fontu (*@FONTSTYLE*, např. tučné nebo kurzivní písmo, horní index apod.).

Element **<Tags>**, zavedený ve verzi 2.1, umožňuje definovat libovolné množství elementů **<LayoutTag>**, **<StructureTag>**, **<RoleTag>**, **<NamedEntityTag>** a **<OtherTag>**, které mohou nést podrobnější informaci o roli nebo významu rozpoznaných částí na stránce. Tímto způsobem lze např. identifikovat razítka, reklamy (**<LayoutTag>**), nadpisy, živé záhlaví (**<StructureTag>**), autora, vydavatele (**<RoleTag>**), jména osob, názvy institucí (**<NamedEntityTag>**), popř. další zaznamenání hodný prvek (**<OtherTag>**).

V rámci elementu **<ReadingOrder>** lze zachytit způsob logického uspořádání rozpoznaných prvků na stránce

z hlediska návaznosti textu, které se liší od fyzického uspořádání na stránce, popř. ve vícestránkovém dokumentu (např. v případě novinových článků). Na elementy, které po sobě následují, se odkazuje pomocí identifikátorů v elementech **<OrderedGroup>**, popř. **<UnorderedGroup>**.

Element **<Layout>** obsahuje informace o rozvržení dílčích oblastí na prostoru digitalizované stránky: jednak okraje (**<TopMargin>**, **<LeftMargin>**, **<RightMargin>** a **<BottomMargin>**), na nichž se můžou vyskytovat specifické části textové struktury (např. marginální poznámky, živé záhlaví, apertury apod.), jednak samotná tisková plocha **<PrintSpace>** nesoucí hlavní textový obsah na straně. Z hlediska dalšího rozčlenění můžou uvedené dětské prvky elementu **<Layout>** obsahovat shodnou strukturu, tj. prvky **<Shape>**, **<TextBlock>**, **<Illustration>**, **<GraphicalElement>** a **<ComposedBlock>**.

Element **<Shape>** může být tvořen samostatnými prvky typu polygon (**<Polygon>**), elipsa (**<Ellipse>**) a kruh (**<Circle>**) a může být součástí libovolného prvku na stránce. Prvek **<GraphicalElement>** slouží ke grafickému oddělení jiných textových či grafických bloků na stránce (např. ornament mezi dvěma kapitolami).

Pomocí hodnot atributů lze u elementu **<Illustration>** určit rozměry ilustrace (*@HEIGHT*, *@WIDTH*, *@HPOS*, *@VPOS*), případně její otočení (*@ROTATION*) a vztah s jiným prvkem na stránce, např. popiskem (*@IDNEXT*), ale i typ zobrazeného objektu (*@TYPE*, např. fotografie, mapa, graf apod.) a případně odkaz na soubor, který danou ilustraci obsahuje (*@FILEID*).

Na závěr základního přehledu elementů ze standardu ALTO zbývají prvky, které nesou textovou informaci, a to **<TextBlock>**, který sdílí stejné typografické vlastnosti na úrovni bloku a je tvořen jedním nebo několika samostatnými řádky (**<TextLine>**). Jeho podřazenými prvky jsou **<String>** (text tvořený písmeny a interpunkčními znaménky) a prázdné prvky **<SP>** (text tvořený prázdným prostorem, tj. mezery a tabulátory), případně **<HYP>** (obsahuje spojovník, popř. jiné dělicí znaménko na konci řádku).

Element **<String>** obsahuje rozpoznaný text v atributu *@CONTENT*. Vedle toho je k dispozici také atribut *@SUBS_CONTENT*, jehož obsah tvoří upravený rozpoznaný text, např. text s rozepsanou zkratkou nebo kompletním slovem, které je v předloze rozděleno spojovníkem. Pomocí hodnoty atributu *@SUBS_TYPE* lze specifikovat, o jaký typ úpravy se jedná. Pomocí atributu *@WC* lze zachytit míru pravděpodobnosti (hodnota v rozmezí 0–1), s níž rozpoznávací software text danému úseku přiřadil. Pravděpodobnost správného přiřazení jednotlivých liter se zachycuje v atributu *@CC*: pro každý znak se uvede číselné hodnocení na škále od 0 (jisté) do 9 (nejméně pravděpodobné), čísla se oddělují mezerou. Atribut *@CS* slouží k zachycení faktu, zda došlo k ruční opravě rozpoznaného textu.

V případě potřeby je možné v elementu **<String>** zachytit také alternativní rozpoznané podoby, a to pomocí vnořeného elementu **<ALTERNATIVE>** a jeho atributu *@PURPOSE*, tentokrát však již bez určení míry pravděpodobnosti. Pomocí vnořených elementů **<Glyph>** lze rozpoznaný text dále rozdělit na jednotlivé glyfy (písmena, spřežky) a určit míru pravděpodobnosti jejich přiřazení pomocí atributu *@GC*. Také u této nejmenší textové jednotky lze zachytit alternativní podobu, a to ve vnořeném elementu **<Variant>** (např. pro případy typu „m“ a „rn“).

PAGE XML (PAGE ANALYSIS AND GROUND-TRUTH ELEMENTS XML)

Standard PAGE XML (Pattern Recognition and Image Analysis Research Lab, 2019) od roku 2008 vyvíjí a spravuje Pattern Recognition & Image Analysis (zkráceně PRImA) Research Lab ze Salfordské univerzity. Jedná se vlastně o skupinu formátů zaměřených na různé fáze procesu rozpoznávání textu. Kromě samotného zachycení textových a grafických prvků na stránce lze tímto standardem popsat také provedené úpravy obrázku a vyhodnocení výsledků rozpoznání (profily, výsledky evaluace apod.).

Aktuální schémata jsou dostupná z webových stránek výzkumné laboratoře PRImA nebo z jejího repozitáře na GitHubu. Schémata se verzují pouze změnou jmenného prostoru, který obsahuje aktuální datum poslední změny, která má dopad na zpětnou kompatibilitu (např. „http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15“). Jednotlivé verze jsou dostupné z úložiště na GitHubu, ty však obsahují kompletní zdrojové kódy k danému datu, nejen samotná schémata. Na stránkách výzkumné laboratoře jsou rovněž dostupné softwarové nástroje, které s tímto standardem pracují, některé z nich lze nalézt také na GitHubu, ale nikoli všechny.

Po formální stránce dodržuje PAGE XML principy pojmenování elementů (počáteční velké písmeno,

např. **<ReadingOrder>**) a atributů (počáteční malé písmeno, např. *@primaryLanguage*); v obou případech se uplatňuje velbloudí notace. Poslední verze standardu z července 2019 obsahuje 60 elementů a 80 atributů. Schéma pro úpravu deformací ze srpna 2014 čítá 11 elementů a 7 atributů. Schéma pro vyhodnocování úspěšnosti OCR z července 2019 obsahuje 71 elementů a 70 atributů.

Dokument PAGE XML, který zachycuje digitální předlohu, tvoří kořenový prvek **<PcGts>** (Page Content – Ground Truth and Storage), který obsahuje dvě základní části: metadata (**<Metadata>**) a samotný obsah stránky (**<Page>**). Metadata obsahují informace o tvůrci dokumentu PAGE XML (**<Creator>**), jeho vzniku (**<Created>**) a úpravách (**<LastChange>**). Existuje zde také prostor pro komentáře (**<Comments>**) a uživatelem definované atributy (**<UserDefined>**/**<UserAttribute>**), tvořené názvem, popisem, typem a hodnotou. Stejné atributy mají i elementy **<MetadataItem>**, které navíc umožňují zaznamenat datum přiřazení a sémantické popisky (**<Labels>**/**<Label>**), které mohou odkazovat na externí zdroje (modely, ontologie apod.).

Mezi hlavní prvky pro popis obsahu stránky patří následující:
Uživatelem definované atributy (**<UserDefined>**/**<UserAttribute>**) a sémantické popisky (**<Labels>**/**<Label>**), které lze použít také na úrovni stránky.

Element **<AlternativeImage>** slouží k zachycení informace o jiném zdroji obrazových dat, např. v černobílém rozlišení. K vymezení okrajů samotné stránky slouží element **<Border>**, což je vhodné pro případy, kdy digitální obrázek obsahuje i jiné prvky typu skenovací pult a podložka. Součástí prvku **<PrintSpace>** jsou údaje o těch částech skenu, které obsahují textový obsah, a to pomocí vymezení koordinát polygonu.

Element **<ReadingOrder>** podobně jako v případě standardu ALTO obsahuje logické uspořádání rozpoznaných prvků na stránce z hlediska návaznosti textu pomocí elementů **<OrderedGroup>** a **<UnorderedGroup>**. Na rozdíl od Alta lze pomocí atributů těchto elementů nebo uživatelem definovaných atributů (**<UserDefined>**/**<UserAttribute>**) a sémantických popisků (**<Labels>**/**<Label>**) tyto části podrobněji analyzovat.

Prvek **<Layers>** umožňuje zachytit případy, kdy stránka obsahuje několik překrývajících se vrstev, přičemž výše položená vrstva překrývá prvky z vrstev, které se nacházejí pod ní.

Součástí prvku **<Relations>**/**<Relation>** je informace o vztahu mezi libovolnými prvky v textu, např. mezi obrázkem a jeho popiskem nebo mezi částmi slova rozdělenými koncem řádku.

Element **<TextStyle>** definuje pomocí samostatných atributů typografické vlastnosti textu, kterých lze ve srovnání s Altem zachytit daleko více, např. barvu pozadí) (*@bgColour* a *@bgColourRgb*), kontrastní barvu vůči pozadí (*@reverseVideo*), styl podtržení (*@underlineStyle*) nebo prokládané písmo (*@letterSpaced*). Některé vlastnosti řeší ALTO pomocí kombinace hodnot v atributu *@STYLE*, zatímco PAGE XML pro ně používá samostatné atributy: tučné písmo (*@bold*), kurziva (*@italic*), horní (*@subscript*) a dolní (*@superscript*) index, přeškrtnutý text (*@strikethrough*) a kapitálky (*@smallCaps*).

Dále může strana obsahovat libovolný počet prvků, které reprezentují specifické části výstavby textu. Jedná se o oblasti, které obsahují: text (**<TextRegion>**), obrázek (**<ImageRegion>**), čáry (**<LineDrawingRegion>**), grafiku (**<GraphicRegion>**), tabulku (**<TableRegion>**), graf (**<ChartRegion>**), mapu (**<MapRegion>**), oddělovač (**<SeparatorRegion>**), matematický (**<MathsRegion>**) a chemický (**<ChemRegion>**) vzorec, notový zápis (**<MusicRegion>**), reklamu (**<AdvertRegion>**), šum (**<NoiseRegion>**), neznámý (**<UnknownRegion>**) a uživatelem definovaný (**<CustomRegion>**) typ oblasti. Většina z nich pomocí atributů specifikuje obvyklé vlastnosti oblasti, jako je rotace (*@orientation*), barva pozadí (*@bgColour*) či informace, zda je součástí prvku další text (*@embText*). Další atributy bývají specifické pro konkrétní typ oblasti, např. počet řádků (*@rows*) a sloupců (*@columns*) u tabulky. Uvedené oblasti se mohou navíc libovolně do sebe vnořovat (např. obrázek jako součást tabulky).

Element **<TextRegion>** obsahuje text v určité oblasti stránky, kterou je možné dále rozdělit na řádky (**<TextLine>**) a následně slova (**<Word>**) a glyfy (**<Glyph>**). Na všech uvedených úrovních jsou k dispozici stejné elementy **<TextStyle>** a libovolný počet prvků **<TextEquiv>**, který reprezentuje kompletní text oblasti (tj. celé oblasti, řádku, slova, nebo glyfu) zachycený jako řetězec znaků v kódování ASCII (**<PlainText>**) nebo Unicode (**<Unicode>**).

ABBYY FINEREADER XML

Společnost ABBYY vyvíjí aplikace pro optické rozpoznání textu od roku 1993. Ve verzích určených pro serverové nasazení nabízí export do formátu XML (ABBYY, 2023), který používá vlastní sadu a strukturu

elementů. Od ostatních standardů se odlišuje mj. větší mírou údajů o procesu a výsledku rozpoznání, které se ukládají (např. zda rozpoznaný výraz pocházel ze slovníku). Zároveň lze při generování dokumentu s výsledky OCR stanovit, jaká míra informací se v něm uloží: bez detailů na úrovni znaku a s různou mírou detailů pro jednotlivé znaky, popř. slova (viz dále).

Základní informace o používaném formátu XML jsou k dispozici na webových stránkách nápovědy ke dvěma různým produktům: ABBYY FineReader Engine a ABBYY Cloud OCR. Další nápověda může být součástí instalace samotného produktu. Informace o jednotlivých verzích a jejich změnách nejsou na stránkách výrobce dohledatelné, není také zřejmé, k jakému datu ta či ona verze vstoupila v platnost. Schémata dokumentů nemají předem určené úložiště, naštěstí se uložení schématu shoduje s názvem jmenného prostoru určité verze. Pojmenování verzí v sobě obsahuje číslo produktu a označení podřazené verze schématu, např. „FineReader10-schema-v1.xml“.

Kořenovým prvkem standardu ABBYY FineReader XML je jeden dokument (**<document>**), který může být tvořen metadaty (**<documentData>**) a libovolným počtem stran (**<page>**).

Samotný dokument může v atributech ukládat údaje o verzi dokumentu (*@version*), programu, který jej vytvořil (*@producer*), počtu stran (*@pagesCount*) a o hlavním (*@mainLanguage*) a dalších jazycích (*@languages*), které se v rozpoznaném textu vyskytují.

Metadata definují formátování odstavců (**<paragraph Styles>**/**<paragraphStyle>**). Jednak popisují jejich vlastnosti coby bloku textu: název (*@name*), role v dokumentu (*@role*), úroveň např. nadpisu, číslování či odsazení (*@roleLevel*), zarovnání (*@align*), mezera před (*@before*) a za odstavcem (*@after*), odsazení prvního řádku (*@startIndent*), odsazení celého odstavce zleva (*@leftIndent*) a zprava (*@rightIndent*), a různé kvality řádkování (*@lineSpacing*, *@lineSpacingRatio*, *@fixedLineSpacing*), jednak určují vlastnosti použitého fontu: výchozí font (*@baseFont*), kurzivní (*@italic*), tučné (*@bold*), podtržené (*@underline*) a přeškrtnuté (*@strikeout*) písmo, kapitálky (*@smallcaps*), měřítko (*@scaling*), proložení (*@spacing*), barvu písma (*@color*) a podkladu (*@backgroundColor*), název (rodina) písma (*@ff*) a jeho velikost (*@fs*).

Další prvek metadat (**<sections>**/**<section>**) obsahuje jednotlivé oddíly vyčleněné v dokumentu, jejichž součástí může být několik toků textu či objektů (**<stream>**), které propojují textové prvky a bloky, v nichž se vyskytuje související text; může jít o text, poznámky pod čarou, plovoucí text nebo nežádoucí části (obsahující nečistoty nebo šum). U každého toku se může uvést číslo počáteční a koncové strany a počet sloupců, na nichž se text vyskytuje (**<mainText>**).

Jednotlivé strany (**<page>**) jsou určené šířkou (*@width*), výškou (*@height*), rozlišením (*@resolution*) a zda jsou koordináty relativní vůči originálnímu obrázku nebo upravenému, např. pootočenému (*@originalCoords*). Dále je strana rozdělena na bloky (**<block>**), které můžou být různého typu (*@blockType*): text, tabulka, obrázek, čárový kód, oddělovač a skupina oddělovačů. Dílčí bloky lze pojmenovat (*@name*), určit jejich umístění na stránce (*@l*, *@t*, *@r*, *@b*, tj. pozice levého, horního, pravého a spodního hraničního bodu). V bloku může být text (**<text>**), řádek tabulky (**<row>**), oddělovací prvky seskupené (**<separatorsBox>**) nebo samostatné (**<separator>**), případně oblast (**<region>**) tvořená obdélníky (**<rectangle>**).

Textový blok je dále tvořen odstavci (**<par>**), odstavce jednotlivými řádky (**<line>**) a ty úseky s identickým formátováním (**<formatting>**). Obsahem posledního elementu může být rozpoznaný text (řetězec znaků) nebo množina jednotlivých znaků (**<charParams>**), přičemž za samostatné znaky se považuje i interpunkce a mezera. Součástí řádku může být variantní rozpoznání (**<wordRecVariants>**/**<wordRecVariant>**) následujících znaků.

Na úrovni odstavce (**<par>**) je prostor pro zachycení iniciály (*@dropCapCharsCount*, *@dropCap-l*, *@dropCap-t*, *@dropCap-r*, *@dropCap-b*), zarovnání (*@align*), odsazení z levé (*@leftIndent*) a pravé (*@rightIndent*) strany, odsazení prvního řádku (*@startIndent*), řádkování (*@lineSpacing*), odstavcový styl (*@style*) a zda odstavec začíná (*@hasOverflowedHead*) nebo pokračuje (*@hasOverflowedTail*) na jiné stránce. I když nejnovější verze ABBYY FineReader Engine poslední dva atrbuty nepodporuje, přesto ve specifikaci zůstávají.

Pro jednotlivé znaky je možné podchytit následující údaje: oblast na stránce (*@l*, *@t*, *@r*, *@b*), míra pravděpodobnosti správného rozpoznání (*@charConfidence*), míra pravděpodobnosti patkového písma (*@serifProbability*), nesoulad znaků v jednom slově (*@suspicious*), výška znaku

(*@characterHeight*), případně nejistá výška znaku (*@hasUncertainHeight*).

Také u jednotlivých slov lze zaznamenat podrobnější údaje z procesu rozpoznávání: zda bylo slovo nalezeno ve slovníku (*@wordFromDictionary*), zda se jedná o číslo (*@wordNumeric*), případně identifikátor (*@wordIdentifier*, tj. např. zkratku, URL apod.).

hOCR (HTML OCR)

Odlišnou cestou než předchozí standardy se ubírá projekt hOCR (HOCR, 2020), který pro zachycení výsledků rozpoznání textu využívá jako „nosič" formát HTML, resp. XHTML. Využívá toho, že většinu strukturních prvků, jako jsou nadpisy, tabulky nebo obrázky, tyto standardy již obsahují. Informace, které pocházejí ze systému pro optické rozpoznání znaků, uchovává pomocí obecně dostupných atributů *@class*, *@style* a *@title*.

Standard hOCR používá sémantické verzování, zdrojové texty specifikace jsou k dispozici v úložišti GitHub v samostatných složkách (https://github.com/kba/hocr-spec), poslední verze také jako samostatná webová stránka. Správa a publikování jednotlivých verzí na GitHubu bohužel nekoresponduje s označením a datací verzí tohoto standardu. Poslední specifikace je dostupná pod licencí CC0 1.0 DEED (https://creativecommons.org/publicdomain/zero/1.0/) a Open Web Foundation Contributor License Agreements (https://www.openwebfoundation.org/the-agreements/the-owf-1-0-agreements-granted-claims/owfa-1-0).

Metadata k rozpoznanému dokumentu, která generuje aplikace pro OCR, se ukládají v hlavičce souboru (X)HTML v elementech **<meta>**. V atributu *@name* je možné specifikovat následující údaje: aplikace včetně verze, která dokument vytvořila (*ocr-system*), seznam možností, které aplikace podporuje (*ocr-capabilities*), počet stran v dokumentu (*ocr-number-of-pages*), seznam rozpoznaných jazyků podle standardu ISO 639-1 (*ocr-langs*), seznam rozpoznaných písem podle standardu ISO 15924 (*ocr-scripts*).

Možnosti podporované aplikací pro OCR (tj. metadata typu *ocr-capabilities*) mohou nabývat následujících hodnot: *ocrp_lang* (použití atributu *@lang* pro označení jazyka), *ocrp_dir* (použití atributu *@dir* pro označení směru čtení textu), *ocrp_poly* (ohraničení oblasti, která nemá tvar obdélníku), *ocrp_font* (standardní vlastnosti použitého fontu), *ocrp_nlp* (míra pravděpodobnosti správnosti rozpoznání), *ocr_embeddedformat_***<formatname>** (seznam vložených formátů, které program umí generovat), *ocr_***<tag>***_unordered* (seznam značek, pro něž nelze určit pořadí čtení).

Pokud jde o rozpoznaný obsah, využívá hOCR atribut *@class* pro definici elementu OCR, který je navázán na daný prvek HTML, obvykle ** ** nebo ****. Atribut musí obsahovat jednu hodnotu, jejíž jméno začíná textem „ocr_", popř. „ocrx_", které samotný standard označuje jako prvky a uvádí je ve špičatých závorkách, jako by se jednalo o prvky XML. Standard rozlišuje několik typů elementů, u některých doporučuje konkrétní elementy HTML, na něž by se měly vázat, což v dalším textu naznačujeme vlnovkou (~). Za podstatné skupinky prvků považujeme následující:

- typografické prvky: strana (**<ocr_page>**), ohraničená oblast (**<ocr_carea>**), řádek (**<ocr_line>**), oddělovač (**<ocr_separator>**), rušivý element (**<ocr_noise>**);
- plovoucí prvky: obecný plovoucí prvek (**<ocr_float>**), text (**<ocr_textfloat>**), obrázek s textem (**<ocr_textimage>**), obrázek (**<ocr_image>**, popř. **<ocr_photo>**; rozdíl mezi těmito prvky není z dokumentace zřejmý), čárová grafika (**<ocr_linedrawing>**), záhlaví (**<ocr_header>**), zápatí (**<ocr_footer>**), číslo strany (**<ocr_pageno>**) a tabulka (**<ocr_table>**);
- logické prvky: dokument (**<ocr_document>** ~ ** **), titul (**<ocr_title>** ~ **<h1>**), autor (**<ocr_author>**),abstrakt (**<ocr_abstract>**), část (**<ocr_part>** ~- **<h1>**), kapitola (**<ocr_chapter>** ~ **<h1>**), oddíl (**<ocr_section>** ~ **<h2>**), pododdíl (**<ocr_subsection>** ~**<h3>**), podpododdíl (**<ocr_subsubsection>** ~ **<h4>**), zobrazení (**<ocr_display>**), citát (**<ocr_blockquote>** ~ **<blockquote>**), odstavec (**<ocr_par>** ~ **<p>**), lineární prvek (**<ocr_linear>**) a popisek (**<ocr_caption>**);
- řádkové prvky: nerozpoznané glyfy reprezentované obrázkem (tj. elementem HTML **<img>**), buď pro jeden (**<ocr_glyph>**), nebo více (**<ocr_glyphs>**) znaků, iniciála (**<ocr_dropcap>**), matematický (**<ocr_math>**) a chemický (**<ocr_chem>**) vzorec, nespecifikovaný řádkový prvek (**<ocr_cinfo>**);
- abstraktní prvky generované při rozpoznávání, které nemají oporu v sazbě nebo logickém členění (a zároveň pro ně nelze využít některý z dříve uvedených prvků): blok textu (**<ocrx_block>**), řádek (**<ocrx_line>**) a slovo (**<ocrx_word>**).

Pokud je potřeba u prvků hOCR (identifikovaných hodnotou atributu *@class*) uvádět parametry, použije se

k tomu hodnota atributu *@title*: parametry začínají stanoveným názvem, následují hodnoty oddělené mezerou (případně seskupené pomocí specifických znaků, např. čárek); další parametry jsou odděleny středníkem, např.
****hello****

hOCR definuje následující parametry: účaří (*baseline*), ohraničující rámeček (*bbox*), tok částí mezi oblastmi, zejména mezi více prvky **<ocr_carea>** nebo mezi **<ocr_carea>** a **<ocr_linear>** (*cflow*), ořezy pro jednotlivá písmena (*cuts*), pevné zakončení řádku (*hardbreak*), cesta k obrázku (*image*), hašová hodnota souboru (*imagemd5*), uvedené číslo strany (*lpageno*), pořadové číslo strany (*ppageno*), míra pravděpodobnosti správnosti rozpoznání (*nlp*), pořadí prvku toku čtení, pokud jej nelze odvodit ze sousledností elementů XHTML (*order*), ohraničení oblasti, která nemá tvar obdélníku (*poly*), rozlišení skenování (*scan_res*), rotace textu ve stupních (*textangle*), ohraničující rámeček jednotlivých glyfů (*x_bboxes*), jméno fontu přiřazené aplikací OCR (*x_font*), velikost fontu přiřazená aplikací OCR (*x_fsize*), míra pravděpodobnosti správného rozpoznání znaku, kterou stanovila aplikace OCR (*x_confs*), identifikace skeneru (*x_scanner*), identifikace zdrojových dat (*x_source*) a míra pravděpodobnosti správného rozpoznání textového řetězce, kterou stanovila aplikace OCR (*x_wconf*).

Kromě výše uvedených prvků a jejich parametrů je možné a v některých případech nutné při tvorbě dokumentu podle standardu hOCR aplikovat i další postupy a pravidla. Například povinné použití elementů **<sup>** a **<sub>** pro horní a dolní index, elementů **<ul>**, **<ol>** a **<dl>** pro seznamy nebo kombinace elementů **<ins>** a **<del>** v případě alternativní interpretace rozpoznaného textu.

STANDARDY NA BÁZI FORMÁTU JSON

Vedle standardů na bázi značkovacího jazyka XML se objevují také formáty, které využívají formát JSON, a to zejména u rozpoznání textu v rámci webových služeb, jako je Cloud Vision API firmy Google nebo Azure AI Services od společnosti Microsoft.

I když se formáty JSON a XML liší vnitřní strukturou, v rámci jednotného značení rozpoznávaných částí budeme i nadále graficky rozlišovat mezi elementy (v JSONu vlastnosti, které mohou obsahovat další vlastnosti, popř. objekty, pole) a atributy (ve formátu JSON vlastnosti, které mohou obsahovat základní hodnotu, tj. text, číslo, popř. hodnoty true, false a null).

CLOUD VISION API

Společnost Google v rámci svých cloudových služeb nabízí vývojové prostředí Vertex AI Vision určené pro zpracování obrazového materiálu. V jeho rámci jsou prostřednictvím rozhraní Cloud Vision API (Google, bez data) dostupné natrénované modely strojového učení pro konkrétní úlohy, jako je rozpoznání objektů nebo textu na obrázku. Programové rozhraní REST API vrací odpovědi ve strojově čitelném formátu JSON. Výstupy z rozpoznání textu jsou součástí objektu typu TextAnnotation, který obsahuje další, hierarchicky uspořádané vlastnosti, které jsou pojmenovány v množném čísle (např. **pages**, **blocks**, **paragraphs**) a obsahují pole podřazených objektů daného typu (stránka, blok, odstavec), případně další vlastnosti vztažené na daný objekt.

Součástí pole **<pages>** je vlastnost *@text*, která obsahuje neformátovaný rozpoznaný text všech analyzovaných stran zalomený do řádků, tj. včetně spojovníků. Dále zde najdeme pole objektů **<blocks>** různých typů (např. „TEXT“, „TABLE“, „PICTURE“ apod.). V textovém bloku se nachází pole odstavců (**<paragraphs>**), které jsou rozčleněny na slova (**<words>**) a následně na pole jednotlivých znaků (**<symbols>**). Samotné znaky jsou součástí vlastnosti *@text*. U všech výše uvedených prvků se můžou objevit shodné vlastnosti: *@confidence*, která v rozmezí od 0 do 1 vyjadřuje důvěryhodnost rozpoznání pomocí OCR v rámci daného objektu; **<boundingBox>** se souřadnicemi vrcholu ohraničujícího rámečku; **<property>** s identifikovaným předělem (**<detectedBreak>**/*@type* a *@isPrefix*) a s polem rozpoznaných jazyků (**<detectedLanguages>**/*@languageCode* a *@confidence*).

OCR FOR IMAGES

Společnost Microsoft nabízí sadu předpřipravených a přizpůsobitelných nástrojů umělé inteligence, rozhraní API a natrénovaných modelů nazvanou Azure AI Služby. Uživatelé mají v rámci balíku služeb pojmenovaného Azure AI Vision (https://azure.microsoft.com/cs-cz/products/ai-services/ai-vision/) k dispozici několik samostatných služeb, případně jejich verzí, které zajišťují rozpoznání textu v rámci obrázku, popř. dokumentu. Jedna z takových služeb, nebo spíš natrénovaných modelů, se v poslední verzi zkráceně nazývá OCR Read (Microsoft, 2023).

Rozpoznaný text je součástí vlastnosti **<readResult>**. Tento objekt tvoří vlastnost *@content*, jež obsahuje

neformátovaný rozpoznaný text všech analyzovaných stran zalomený do řádků, tj. včetně spojovníků, a pole stránek (**<pages>**). Ke každé stránce je k dispozici pole slov (**<words>**) a pole řádků (**<lines>**), které jsou (z hlediska struktury dokumentu JSON) na sobě nezávislé. Jednotlivé elementy těchto polí mají sadu shodných vlastností, popř. objektů: prostý text daného prvku (*@content*), pozici v toku textu (****), tj. pozice od začátku stránky (*@offset*) a délka textu (*@length*) a umístění ohraničujícího rámečku (**<boundingBox>**). U slov se uvádí důvěryhodnost rozpoznání pomocí OCR (*@confidence*, v rozsahu od 0 do 1). Za slovo se považuje text ohraničený mezerami, resp. prázdným místem, takže jejich součástí jsou i interpunkční znaménka. Na rozdíl od Cloud Vision API se spojovník žádným specifickým způsobem neoznačuje, bývá součástí rozpoznaného textu jako znak pro minus (U+002D).

SROVNÁNÍ STANDARDŮ

Standardy na bázi formátu JSON nabízejí ve srovnání se standardy na základě XML velmi omezený repertoár údajů o jednotlivých rozpoznaných částech. Výstup od firmy Microsoft např. neobsahuje informaci o rozpoznávaném jazyku ani o rozdělení slov a text není rozčleněn do odstavců, pouze do řádků a slov, tj. nedostává se ani na úroveň jednotlivých znaků. I když uvedené údaje jsou k dispozici ve výstupech od firmy Google, u obou poskytovatelů zcela chybí informace o typografických kvalitách rozpoznaného textu. Cloud Vision API identifikuje sice spojovníky, ale slova takto rozdělená je nutné spojit do jednoho na základě analýzy dílčích slov a jejich obsahu. Je-li potřeba pracovat s prostým textem nebo zobrazit konkrétní výraz na stránce, budou výše zmíněné formáty vhodné, ale v případě pokročilejšího využití, například metodami počítačového zpracování přirozeného jazyka (lemmatizace, rozpoznání entit apod.) je vzhledem k zachování spojovníku v prostém textu nutná jejich další úprava. S ohledem na rozčlenění do odstavců se jako vhodnější pro tyto případy jeví výstup firmy Google. Pro náročnější úkoly, jako je kupříkladu příprava elektronické edice, od níž se očekává co nejvěrnější zachycení předlohy, včetně některých typografických kvalit, se oba uvedené nástroje pro OCR rozhodně nehodí.

V dalším popisu se zaměříme na srovnání standardů, které pro zprostředkování údajů o rozpoznání textu využívají formát XML. Důležitými hledisky budou kvalita jejich zpracování, podpora uživatelů a zejména možnosti rekonstrukce původního textu. Zaměříme se zejména na standardy ALTO (používá se v digitálních knihovnách Kramerius), PAGE XML (novější a velmi komplexní standard, využívaný mj. v projektu PERO OCR) a ABBYY FineReader XML (spravuje jej společnost, která vyvíjí software pro rozpoznání textu).

DOKUMENTACE A PODPORA

Z hlediska dokumentace má nejslabší podporu formát ABBY FineReader XML. Veřejně dostupná dokumentace čítá dvě webové stránky: jedna velmi stručně obecně přibližuje obsah vygenerovaného dokumentu XML, druhá uvádí význam používaných elementů a atributů, které se však můžou lišit od uváděných příkladů. O něco lépe je na tom formát PAGE XML, jehož dokumentace je součástí samotného dokumentu XSD. Starší verze jsou dohledatelné v repozitáři na GitHubu, ale nejsou zřetelně vydělené (lze je najít v historii zdrojového kódu). Samostatné verze vyčleňuje a zachovává standard hOCR, i když zpracování není jednotné (verze 1.0 existuje jako dokument Google); jeho nedostatkem je rovněž fakt, že dokumentace není kompletní a vzhledem k chybějícím příkladům není vždy jasné, k čemu se mají definované prvky použít. Nejlépe lze v tomto bodě hodnotit standard ALTO, který přehledně odděluje dílčí verze a obsahuje popis hlavních změn mezi jednotlivými verzemi, a to jak na webových stránkách, tak v dokumentu XSD.

Pokud jde o „formální čistotu“ standardu, reprezentuje ALTO nejnižší kvalitu (nejednotné používání velikosti písmen a oddělovačů v názvech elementů a atributů). Všechny ostatní formáty dodržují jednotné principy, i když například hOCR lze vytknout, že se některé používané elementy zdají být synonymní, a tedy zbytečné. U standardu, který zvolila společnost ABBYY, je definice typu dokumentu uložena v souboru s příponou „.xml“, i když se obvykle používá přípona „.xsd“.

VÍCESTRÁNKOVÉ DOKUMENTY A ROZMĚRY

Kořenovým prvkem standardu ALTO je jedna strana (**<alto>**), takže pro zachycení dokumentu tvořeného více stranami je potřeba použít jiný standard, např. METS, který zachová řazení jednotlivých stran za sebou. Podobně je na tom PAGE XML. Standardy ABBY FineReader XML a hOCR umožňují zpracovat digitální předlohu v jednom vícestránkovém dokumentu.

Všechny formáty uvádějí údaje o výšce a šířce naskenovaného obrázku stránky (názvy atributů se samozřejmě liší). Pro umístění rozpoznaných prvků na stránce používá každý standard vlastní způsob: ALTO uvádí u bloků i slov výšku, šířku a horizontální a vertikální posun, který se vztahuje k levému hornímu rohu; PAGE XML používá

podřízený element <Coords>, jehož atribut *@points* obsahuje seznam bodů (série dvojic čísel oddělených čárkou, které jsou od sebe odděleny mezerou), definujících cestu, která prvek ohraničuje; ABBYY FineReader XML používá údaje o pozici levého, horního, pravého a spodního hraničního bodu (*@l*, *@t*, *@r*, *@b*).

JAZYKY

Pro správné rozpoznání textu, ale i jeho následné zpracování je důležitá informace o jazyce, v němž jsou dokument nebo jeho dílčí části vytvořeny.

Standard ALTO zavedl označení jazyka na úrovni textového bloku (**<TextBlock>**), řádku (**<TextLine>**) a slova (**<String>**) pomocí atributu *@LANG*, a to od verze 2.1, od verze 4.4 jej lze použít i celou stranu (**<Page>**). Do verze 2.0 se používal atribute *@language*. Na hodnotu atributu se kladou stejná omezení jako na obecný atribut *@xml:lang*, která vycházejí z doporučení BCP 47 (https://www.rfc-editor.org/info/bcp47). Analýza materiálu z digitálních knihoven Kramerius dokládá použití atributu *@language* pouze u elementu **<TextBlock>**, u nižších jednotek se jazyk nezachycuje.

PAGE XML umožňuje rozlišit mezi hlavním (*@primaryLanguage*) a druhým (*@secondaryLanguage*) jazykem na úrovni stránky (**<Page>**). Hodnoty lze změnit na nižších úrovních, tj. v rámci regionu, řádku (**<TextLine>**) nebo slova (**<Word>**). Názvy by měly být pojmenovány celým anglickým jménem podle standardu z rodiny ISO 639, ale samotný seznam, který je součástí definice PAGE XML, se v některých případech od tohoto standardu odchyluje a navíc zavádí hodnotu pro další, neuvedený jazyk („other“).

ABBY FineReader XML obsahuje na úrovni celého dokumentu nepovinné atributy *@mainLanguage* (převažující jazyk v dokumentu) a *@languages* (všechny identifikované jazyky) a dále na úrovni souvislého úseku se stejným formátováním (**<formatting>**) povinný atribut *@lang*. Hodnota atributu vychází z nějakého vlastního seznamu pojmenování (např. „EnglishUnitedStates“, „GermanStandard“). Seznam hodnot pro jazyky nejspíš zachycuje tato stránka: https://www.simpleindex.com/wiki/index.php?title=FineReader.

FORMÁTOVÁNÍ TEXTU

Všechny hlavní standardy (ALTO, PAGE XML a ABBYY FineReader XML) umožňují definovat styly, tj. opětovně použitelné vlastnosti formátování, a to na úrovni celého dokumentu (ABYY FineReader XML: pro odstavce **<paragraphStyle>** a v rámci jednoho odstavce pak několik znakových stylů, tj. **<fontStyle>**) nebo jednotlivé stránky (ALTO: **<ParagraphStyle>** a **<TextStyle>**, tj. styly pro odstavce a textové řetězce); na takto definované styly lze odkázat u objektů, které dané formátování aplikují. Dostupné ukázky výstupů ve formátu ABBYY FineReader XML sice v elementu **<formatting>** odkazují na definované znakové styly, zároveň uvádějí identické atributy včetně jejich hodnot jako definovaný styl, takže se taková definice stylů zdá redundantní. PAGE XML definuje výchozí znakové formátování na úrovni stránky (**<TextStyle>**): pokud se na konkrétních místech formátování odlišuje, definují se hodnoty, které se od těch výchozích liší, takže se principiálně od předchozích dvou standardů odlišuje. Pro zachycení formátování odstavců používá PAGE XML vybrané atributy elementu **<TextRegion>**.

U jednotlivých standardů se liší repertoár rozpoznávaných vlastností formátování. S minimem údajů operuje ALTO, další formáty nabízejí vlastnosti navíc, přičemž na úrovní odstavců má k dispozici více atributů ABBYY FineReader XML, zatímco PAGE XML umožňuje podchytit více detailů u formátování textového úseku (např. horní a dolní index nebo barvu podkladu).

Jelikož mají v Altu identifikátory stylu podobu jedinečných identifikátorů (GUID) bez vazby na samotné kvality formátování, může mít tentýž formát při zpracování stejného dokumentu v různých instancích programu pro OCR odlišné identifikátory, takže je potřeba vždy vycházet z konkrétních hodnot stylu.

K dalším komplikacím dochází, když software pro rozpoznání textu identifikuje v originále velké množství formátování (např. několik velikostí písma s odstupňováním po 0,5 bodu), byť byl při sazbě použit jenom omezený repertoár písem. Tato rozkolísanost je daná kvalitou originálu i digitální kopie, ale může mít dopad na další automatizované zpracování rozpoznaného textu, které počítá s určitými kvalitami typografického zpracování, jako je tomu např. u slovníků.

UMÍSTĚNÍ A TOK TEXTU

V této části uváděné názvy elementů budou dodržovat následující pořadí standardů: ALTO, PAGE XML, ABBYY FineReader XML. Standardy umožňují s textem jakožto nosičem informace pracovat s různou mírou detailu, shodné je rozdělení na bloky textu (**<TextBlock>**, **<block>**, **<TextRegion>**), definované sdílenými blokovými typografickými vlastnostmi (výška řádku, mezera mezi odstavci apod.), dále pak rozčlenění na samostatné řádky (**<TextLine>**, **<line>**,

<TextLine>), slova (**<String>**, **<Word>**, ∅) a případně jednotlivé znaky (**<Glyph>**, **<charParams>**, **<Glyph>**). ABBYY FineReader XML člení textový blok na samostatné odstavce (**<par>**) a v jejich rámci teprve na řádky. Pro vymezení hranic jednotlivých slov se u něj používá atribut *@wordStart* v rámci prvku **<charParams>**.

U složitějších publikací, jako jsou ilustrované knihy nebo noviny a časopisy, je důležité zachytit zamýšlený tok textu, který bývá přerušen ilustracemi, inzeráty nebo bloky samostatného, nesouvisejícího textu, v některých případech může text navazovat i o několik stránek dále. Z hlediska dalšího automatizovaného zpracování a jazykové analýzy je důležité zachovat myšlenkovou i syntaktickou návaznost rozdělených částí textu. Standard ALTO (od verze 4.3) umožňuje v metadatech ke stránce (**<alto>**/**<ReadingOrder>**) definovat pomocí odkazů na textové bloky jejich správné kontinuální pořadí (**<OrderedGroup>**), stejným způsobem lze zachytit i bloky textu, které spolu nesouvisí (**<UnorderedGroup>**). Obdobný způsob aplikuje PAGE XML, využívá k tomu stejně pojmenované prvky, ale jejich obsah, tj. zachycení vztahu mezi neseřazenými a seřazenými prvky, může být komplikovanější (**<RegionRefIndexed>**, **<OrderedGroupIndexed>**, **<UnorderedGroupIndexed>**). Tento standard obsahuje u textových bloků i atribut *@continuationt*, určující, zde je region nebo skupina pokračováním předchozího prvku, ale nepodařilo se najít jediný příklad jeho použití. ABBYY FineReader XML používá k zachycení posloupnosti textu v dokumentu oddíl metadat s elementy **<sections>** a **<section>**: vytvořením několika toků (**<stream>**) s posloupností odkazovaných elementů (**<elementId>**) lze definovat, jak jdou jednotlivé části po sobě. Element **<stream>** pomocí atributů *@beginPage* a *@endPage* vymezuje začátek a konec kontinuálního textu v dokumentu.

DĚLENÍ SLOV

Oblast vymezená pro zachycení textu hraje důležitou roli již od samých počátků zaznamenávání psaného jazyka. Rozpor mezi délkou myšlenky, vymezeným fyzickým prostorem pro její zachycení a typografickými pravidly vyústil v dělení slov na konci řádku. V současném psaném jazyce se pro tyto účely v latince používá spojovník (tiret, divis), který mohl mít v dřívějších dobách i různou podobu (připomínající např. rovnítko, viz např. https://ocr-d.de/en/gt-guidelines/trans/trSilbentrennung.html). Nerozpoznaný spojovník má dopad na věcnou i gramatickou správnost delšího úseku.

Standard ALTO používá pro zachycení spojovníku samostatný element **<HYP>**, zajímavé přitom je, že má povinný atribut *@CONTENT*, který na základě analýzy dat z digitálních knihoven Kramerius obvykle obsahuje hodnotu „175“, což je desítková hodnota znaku pro logickou negaci (¬), asi v 10 % případů obsahoval atribut přímo znak pro tzv. volitelnou pomlčku (U+00AD) a pouze v těchto případech se u předchozího a následujícího slova (**<String>**) objevil atribut *@SUBS_CONTENT*, který obsahuje kompletní podobu rozděleného slova.

PAGE XML nabízí pro propojení dvou částí rozděleného slova element **<Page>**/**<Relation>** typu „join“, bohužel dohledatelné dokumenty obsahují příklady pouze pro iniciálu a odstavec, nikoli pro rozdělené slovo. Pro poslední verzi standardu se počítalo se dvěma atributy, které měly pro tyto účely sloužit (*@mergeWithNextRule* a *@mergeWithNextRuleData*): zůstaly však zakomentované a do standardu se nedostaly. Rozlišení mezi spojovníkem a pomlčkou je tedy na uživateli, popř. softwaru, který s dokumentem pracuje.

ABBYY FineReader XML rovněž nenabízí žádný prostředek pro speciální zacházení s rozděleným slovem na konci řádku. V ukázkách, které jsme měli k dispozici, byl spojovník rozpoznán jako znak pro logickou negaci (¬).

Zmiňované tři standardy se odlišují ještě v jedné věci, která se týká textu. Zatímco ALTO za nejrozsáhlejší jednotku, u níž může být uveden rozpoznaný, popř. alternativní text, považuje slovo (**<String>**), PAGE XML může díky podřízenému elementu **<TextEquiv>** uvádět text i u rozsáhlejších jednotek na stránce, tj. pro textový blok, řádek, slovo a jednotlivý znak (viz výše). ABBYY FineReader XML umí také uchovávat text větších celků, než je znak nebo slovo, ale na rozdíl od PAGE XML neumí tyto úrovně kombinovat: uživatel si musí zvolit pouze jednu z nich. Konkrétně je možné vygenerovat kontinuální text na úrovni řádku pro úseky se stejným formátováním.

VYUŽITÍ STANDARDŮ NA BÁZI XML

EXPORTNÍ FORMÁT APLIKACÍ PRO OCR

Všechny představené formáty lze využít pro uchování informace o zdrojových datech a rozpoznaných prvcích včetně textu. Export do uvedených formátů je součástí nástrojů pro rozpoznání textu, např. ABBYY FineReader Engine (ALTO 3.1), docWizz (ALTO), kraken (ALTO, PAGE XML), OCR4all (PAGE XML 2019-07-15), tesseract (hOCR, ALTO 4.1). Využívají je také projekty určené pro rozpoznání textu pomocí strojového učení, např. Transkribus (PAGE XML 2013-07-15, ALTO 2.0 + ALTO 4.2), eScriptorium (PAGE XML 2019-07-15, ALTO 4.2) nebo PERO OCR (PAGE XML 2019-07-15, ALTO 2.0).

NOSIČ INFORMACE V DIGITÁLNÍCH KNIHOVNÁCH

Samozřejmě lze tyto formáty využít pro archivaci a následně prezentaci rozpoznaných digitalizovaných dokumentů. Digitální knihovna Trove Australské Národní knihovny (https://trove.nla.gov.au) například nechala digitalizovat periodika externím dodavatelem, který dodal data ve formátu METS a ALTO. Bavorská zemská knihovna v rámci svých digitálních sbírek (https://dafo.digitale-sammlungen.de) využívá standard hOCR pro zobrazení textové vrstvy k digitalizované stránce. Data ve formátu hOCR z uživatelského rozhraní přímo dostupná nejsou, ale lze je získat pomocí programového rozhraní REST API. Formát ALTO je součástí standardů pro metadata v Národní digitální knihovně (https://standardy.ndk.cz/ndk/standardy-digitalizace), a to ve verzi 2.0 (pro tištěné dokumenty) a 3.0 (pro zvukové dokumenty). Proto jej nabízejí české digitální knihovny využívající software Kramerius. Rozpoznaná stránka je dostupná jednak z uživatelského rozhraní, jednak programově pomocí programového rozhraní REST API. Dalším úložištěm digitálních publikací, které nabízí rozpoznaný text v několika formátech, je Internet Archive (https://archive.org), který funguje jako agregátor dat od různých institucí, přičemž se nezaměřuje pouze na tištěná a rukopisná díla, ale i další autorská díla, jako jsou webové stránky, audio, video, software nebo obrázky. Tištěná produkce bývá dostupná ve formátu ABBYY FineReader XML nebo hOCR.

ZMĚNY VE STANDARDU ALTO

Jak již bylo uvedeno výše, ALTO ve verzi 2.0 je součástí standardů pro metadata v Národní digitální knihovně, a používá se tedy i v českých digitálních knihovnách Kramerius. Používaná verze pochází z roku 2014, přičemž během téměř deseti let se do standardu ALTO, aktuálně ve verzi 4.4, promítly změny, které doplňují nebo zpřesňují informace o zpracování stránky i textu: k dispozici je obecnější element **<Processing>** vedle (stále platného) **<OCRProcessing>**; atribut *@ROTATION* definuje výchozí úhel otočení textu, případně ilustrací na celé stránce; v předchozích verzích bylo možné definovat otočení jenom na úrovni bloku. V případě textu umožňuje zavedení prvku **<ReadingOrder>** definovat jeho tok mezi dílčímu bloky na stránce, elementy **<Glyph>** slouží pro identifikaci jednotlivých rozpoznaných znaků, obrazce typu **<Shape>** mohou být součástí řádku a přibily i nové vlastnosti formátování (hodnota pro přeškrtnuté písmo v rámci *@FONTSTYLE*). Díky atributům *@LANG* a *@OTHERLANGS* je možné zachytit rozpoznané jazyky na stránce. Pro cizojazyčné publikace je důležitý atribut *@BASEDIRECTION*, který určuje směr čtení textu (zleva doprava, shora dolů apod.).

ZÁVĚR

Standardy pro zachycení informací o rozpoznaném obsahu digitalizovaných dokumentů významnou měrou přispívají k uchování a šíření kulturního dědictví. Díky své povaze nositele základní pravdy („ground truth“) slouží také k vylepšování dosavadních postupů při rozpoznání tištěného i ručně psaného textu. Významnou roli hraje v digitálních knihovnách, jimž umožňuje nabídnout čtenáři nejen digitální kopii originálního dokumentu se všemi přednostmi i chybami, ale i elektronicky čitelnou a strojově zpracovatelnou verzi. Z mnoha formátů, které byly a jsou k dispozici, jsme se zaměřili na standardy využívající značkovací jazyk XML, byť jsme popsali i výsledky rozpoznání textu pomocí cloudových služeb, které k přenosu informací využívají formát JSON. Do popředí našeho zájmu se dostaly zejména standardy ALTO a PAGE XML, které slouží jako výstupní nebo vstupní formát moderních nástrojů pro OCR a HTR. Formát ABBYY FineReader XML je důležitý mj. proto, že za jeho vývojem stojí firma, která vyvíjí OCR software pro zpracování velkého objemu dat.

Srovnání těchto standardů z několika hledisek ukázalo na jejich slabá i silná místa: kvalita, dokumentace, podpora, propracované a opomíjené části rozpoznaného dokumentu. Viděli jsme i to, jak se vzájemně ovlivňují (např. zachycení toku textu nebo rozpoznaného jazyka). Podstatná část srovnání se věnovala způsobům zachycení kontinuálního textu, a to nejen na jedné stránce, ale i přes jejich hranici. Pro rozsáhlé texty se více hodí standard ABBYY FineReader XML, který obsahuje elementy pro vícestránkový dokument, zatímco zbývající dva formáty musejí být součástí dokumentu se samostatnou strukturou, např. METS, ale ani to neřeší provázání prvků mezi dvěma stranami. S rozdělenými slovy na úrovni řádku umí zase lépe pracovat formát ALTO, který je z tohoto pohledu vhodný doplněk k digitálnímu obrazu, zejména pokud se použije jeho novější verze s informacemi o otočení stránky, detekovanými jazyky apod. Některé standardy jsou připravené na integraci s nástroji pro počítačové zpracování přirozeného jazyka a umožňují obohatit rozpoznaný text o další údaje, např. identifikaci rozpoznaných entit (ALTO) nebo o libovolné údaje pomocí popisku a uživatelem definovaných vlastností (PAGE XML). Jiné umožňují do velké míry zachytit samotný proces rozpoznání textu (ABBYY FineReader XML). Z tohoto pohledu je rovněž zajímavé, že všechny hlavní standardy používají odlišný způsob identifikace rozpoznaného jazyka.

DEDIKACE

Publikace vznikla v rámci Institucionální podpory na dlouhodobý koncepční rozvoj výzkumné organizace – Moravská zemská knihovna v Brně.

ZDROJE

ABBYY, 2023. XML Schema Description. In: ABBYY. *ABBYY Online Help* [online]. [cit. 2023-11-05]. Dostupné z: https://help.abbyy.com/en-us/finereaderengine/12/user_guide/specifications_xmlschema/

GOOGLE, bez data. Cloud Vision API. In: GOOGLE. *Google Cloud* [online]. [cit. 2023-11-01]. Dostupné z: https://cloud.google.com/vision/docs/reference/rest

HOCR: OCR Workflow and Output embedded in HTML, 2020. In: *OCR-D: DFG-funded Initiative for Optical Character Recognition Development* [online]. [cit. 2023-11-05]. Dostupné z: https://kba.github.io/hocr-spec/1.2/

Kramerius [online], bez data. Praha: Knihovna Akademie věd ČR [cit. 2023-11-01]. Dostupné z: https://system-kramerius.cz

LEHEČKA, Boris, 2023. Standardy pro zachycení výsledků rozpoznání tištěného a rukopisného textu (OCR, HTR). In: *Boris Lehečka* [online]. [cit. 2023-11-05]. Dostupné z: https://github.com/daliboris/ocr-standards

LIBRARY OF CONGRESS, 2023. ALTO: Analyzed Layout and Text Object. In: LIBRARY OF CONGRESS. *Library of Congress* [online]. [cit. 2023-11-05]. Dostupné z: https://www.loc.gov/standards/alto/

LIBRARY OF CONGRESS, 2023. METS: Metadata Encoding & Transmission Standard. In: LIBRARY OF CONGRESS. *Library of Congress* [online]. [cit. 2023-11-05]. Dostupné z: https://www.loc.gov/standards/mets/

MICROSOFT, 2023. Document Intelligence read model. In: MICROSOFT. *Microsoft Learn* [online]. [cit. 2023-11-01]. Dostupné z: https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/concept-read?view=doc-intel-3.1.0

PATTERN RECOGNITION AND IMAGE ANALYSIS RESEARCH LAB, 2019. PAGE XML. In: PATTERN RECOGNITION AND IMAGE ANALYSIS RESEARCH LAB. *PRImA Research Lab* [online]. [cit. 2023-11-05]. Dostupné z: https://github.com/PRImA-Research-Lab/PAGE-XML

PERO OCR: demonstration application [online], bez data. Brno: Vysoké učení technické v Brně [cit. 2023-11-04]. Dostupné z: https://pero-ocr.fit.vutbr.cz

■ Článok bol recenzovaný