Over IMPACT

Hoe kan de toegang tot historische teksten zo verbeterd worden dat ze net zo toegankelijk worden als hun van oorsprong digitale tegenhangers?

Op 1 januari 2008 is het project IMProving ACcess to Text (IMPACT) gestart. Doel van het project is om massadigitalisering en toegankelijkheid van het Europese gedrukte cultureel erfgoed significant te verbeteren, niet alleen door de bestaande OCR-technologie te vernieuwen, maar ook door het ontwikkelen en inzetten van taaltechnologieën om de historische taalbarrière te overbruggen.

Massadigitalisering is de laatste jaren een belangrijk aandachtspunt voor bibliotheken in de hele wereld. Miljoenen pagina's worden gescand. Echter, als bibliotheken naast de afbeelding van een tekst ook de tekst zelf willen kunnen aanbieden, dan lukt dat voor historisch tekstmateriaal met de bestaande OCR-technieken (OCR: optische tekenherkenning) niet en overtikken van deze teksten is te tijdrovend en te kostbaar. Een consortium van vijftien instellingen uit Europa, Israël en Rusland (nationale en universiteitsbibliotheken, onderzoeksinstellingen en bedrijven) heeft zich verenigd om daar iets aan te doen.

Er zal een Best Practiceleidraad gedefinieerd worden m.b.t. de operationele context voor digitalisering. De verschillende binnen IMPACT ontwikkelde technieken zullen 'interoperabel' zijn en er zal ook worden voorzien in een samenhangend programma van verspreiding, trainingen en demonstraties dat gericht is op capaciteitstoename zowel binnen als buiten de deelnemende instituten.

IMPACT streeft ernaar om

1. OCR-software en -technologieën te ontwikkelen die de nauwkeurigheid van de huidige state-of-the-art software substantieel verbeteren, en die het voor het eerst mogelijk zullen maken grote hoeveelheden gedigitaliseerde historische teksten in elektronische tekst om te zetten.
2. Een softwaresysteem te leveren dat de implementatie mogelijk maakt van nieuwe ideeën op het gebied van webgebaseerde collaboratieve correctie.
3. Taaltools en lexica te ontwikkelen om onafhankelijk van de historische varianten van een taal toegang te bieden tot historische teksten.
4. Toepassers van deze tools te steunen zodat meer Europese historische lexica gebouwd kunnen worden.
5. Een aantal kleinere modules te ontwikkelen, zoals toolkits voor beeldverbetering en beeldsegmentatie, functionele parsers etc., met als doel de automatische tekstherkennning en/of toegang tot historisch tekstmateriaal te ondersteunen.

De afdeling Taalbank Nederlands leidt de werkpakketten rondom lexiconbouw en –toepassing ten behoeve van tekstontsluiting, zal een lexicon bouwen voor het Nederlands (GiGaNT) en zal ook verantwoordelijk zijn voor de training dienaangaande. Daarnaast wordt ook meegewerkt aan het ontwikkelen van technieken om de OCR te verbeteren met behulp van taalmodellen en historische lexica.

IMPACT wordt gecoördineerd door de Koninklijke Bibliotheek en gefinancierd binnen het Zevende Kaderprogramma van de Europese Commissie (FP7). De doorlooptijd is vier jaar. Meer informatie op de Impact project website: http://www.impact-project.eu