Knowledge extraction from archives of natural history collections

Stork, L.

Samenvatting

Beschrijvende kennis over de natuurlijke wereld vormt begrip over de verschillende soorten entiteiten die erin leven, hoe ze invloed uitoefenen op en beïnvloed worden door hun veranderende omgeving, en de processen die hun diversiteit teweegbrengen. Dergelijke kennis is cruciaal als het erom gaat beter geïnformeerde beslissingen te nemen voor beleid dat van invloed is op de natuurlijke diversiteit van de wereld, van organismen tot aan ecosystemen. Om kennis over biodiversiteit te vergaren, worden al sinds eeuwen expedities opgezet naar biodiverse plekken om daar levende organismen te beschrijven, illustreren, en verzamelen. Wereldwijd bestaan er vele verzamelde collectieobjecten zoals specimens, veldnotities, soortillustraties, en andere bronnen, opgeslagen in musea en andere instituten. Helaas blijven deze collectieobjecten vaak onderbelicht, vooral vanwege hun complexe contextafhankelijke karakter, het feit dat ze kennis vaak slechts impliciet overdragen, en dat objecten uit verzamelingen vaak verspreid zijn over verschillende collecties en instituten, waardoor hereniging bemoeilijkt wordt en gegevens om deze reden soms incompleet zijn.

Door de hoeveelheid aan historische en hedendaagse collecties samen te brengen op het wereldwijde web (the World Wide Web), als één wereldwijde natuurhistorische collectie, wordt het mogelijk gedetailleerde spatio-temporele analyses te maken van de natuurlijke wereld en van veranderende praktijken in de natuurhistorie. Het bundelen en destilleren van kennis uit grote collecties digitale en fysieke natuurhistorische objecten en het opslaan van het resultaat als gestructureerde, wereldwijd herbruikbare, en toegankelijke kennis, bevordert de samenwerking binnen de onderzoeksgemeenschap en bevordert daarmee het onderzoek en de ontdekking van nieuwe kennis. Het semantisch web (the Semantic Web) biedt een raamwerk om kennis op een zodanige manier op te slaan: *informatie op het wereldwijde web een goed gedefinieerde betekenis geven, waardoor computers en mensen beter kunnen samenwerken aan deze informatie.*[1]

[1] Een geparafraseerd fragment uit een artikel welke gepubliceerd is in de Scientific American van mei 2001, genaamd "Het Semantische Web": "Het Semantische Web is een uitbreiding van het huidige web waarin informatie een duidelijk gedefinieerde betekenis krijgt, waardoor computers en mensen beter kunnen samenwerken (vertaling vanuit het Engels)"

SAMENVATTING

In dit proefschrift analyseren we verschillende methoden om (i) rijke kennis te extraheren die opgeschreven staat in verschillende bronnen van natuurhistorische collecties en (ii) het resultaat op het semantische web te publiceren als machinaal leesbare kennis, zodat anderen deze kennis kunnen hergebruiken voor eigen onderzoek of integratie met eigen collectiegegevens.

Eerst analyseren we systeemontwerpen voor het extraheren van informatie uit documenten met daarin soortobservaties, zoals veldnotities en wetenschappelijke illustraties. Voor de extractie van informatie en kennis uit veldnotities, pleiten we voor een benadering die kwaliteit verkiest boven kwantiteit, rijke semantische annotatie boven volledige teksttranscriptie. Veldnotities zijn een uitdaging om mee te werken vanwege een verscheidenheid aan factoren, zoals de evoluerende visuele stijl van een enkel alfabet en historisch, moeilijk leesbaar handschrift. Semantische annotatie spoort domeinexperts aan om hoge kwaliteit data te produceren en, door de kennis-intensieve aard van de taak zorgt voor intrinsieke motivatie. Verder wordt hierbij ook de minimale hoeveelheid informatie geformaliseerd die nodig is voor voldoende digitalisering.

Allereerst analyseren we welk type semantische informatie domeinexperts gebruiken om collecties te doorzoeken en welke metadata ze nodig hebben om hun collectiegegevens te integreren. Vervolgens analyseren we hoe dergelijke kennis opgeslagen kan worden aan de hand van de principes van Findable, Accessible, Interoperable, and Reusable (FAIR) data, om verder wetenschappelijk discours en ontdekking van nieuwe kennis aan te moedigen. In het Nederlands vertaalt FAIR naar data die vindbaar, toegankelijk, interoperabel, en herbruikbaar opgeslagen moeten worden. Als resultaat van het hierboven beschreven proces beschrijven we een webapplicatie voor de semantische annotatie van natuurhistorische archiefcollecties: de SFB-Annotator. De webapplicatie, met het onderliggende semantische model, wordt verder ontwikkeld binnen het LInking Notes of NAturE (LINNAE)–project.

Daarnaast stellen we een methode voor om het semantische annotatie proces te automatiseren. Het handmatig extraheren van kennis uit archieven is een tijdrovend en arbeidsintensief proces. We laten zien dat we wetenschappelijke namen kunnen identificeren en classificeren in handgeschreven veldnotities, met behulp van sterke aannames gebaseerd op de kennis van domeinexperts over de structuur en inhoud van observatierecords.

Ten slotte analyseren we de extractie van kennis uit wetenschappelijke illustraties. Geautomatiseerde identificatie van soorten is een lastige taak vanwege het feit dat het grootste deel van de kansmassa van soorten zich in de staart van de distributie bevindt en er miljoenen klassen in huidige soortentaxonomiën bestaan. Hierdoor is het lastig om modellen te

creëren die zowel veelvoorkomende als zeldzame soorten kunnen herkennen. We stellen voor om het probleem aan te pakken met zero-shot learning (ZSL). Hoewel openstaande kwesties blijven bestaan—bijv., een verschuiving in distributie tussen collecties van wetenschappelijke illustraties, voortkomend uit verschillen in papiersoorten, illustratiestijl, en granulariteit van afgebeelde objecten—faciliteert ZSL het leren van een model met behulp van achtergrondkennis, welke naar onze mening cruciaal is voor het leren van modellen die automatisch kennis uit kleine datasets met heterogene gegevens kunnen extraheren.