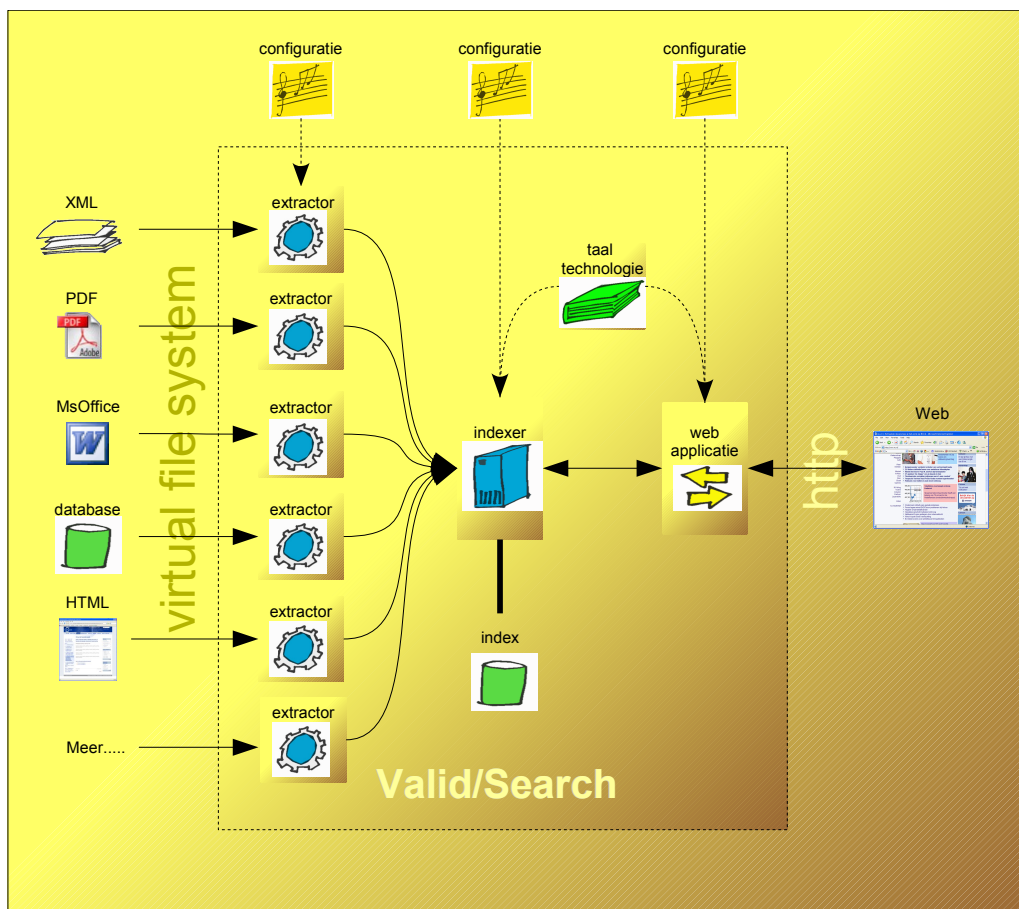


Valid/Search leaflet

Valid/Search is een licentie-vrij softwareproduct dat het bevragen van een veelheid aan gegevensbronnen mogelijk maakt. Het ontsluit onder andere MsOffice, Open Office, PDF, XML, HTML documenten en gegevens in administratieve databases via het web. Zulke gegevens kunnen zich bevinden in verschillende en verspreide (legacy-) systemen. Deze worden samengebracht en benaderd alsof het één systeem betreft. Tenslotte is een auteur een auteur, hoe en waar deze dan ook is vastgelegd...

De volledige tekst van de bronnen wordt bevragd, in combinatie met meer gerichte zoekvragen op beschrijvende gegevens ("metadata") zoals auteur en verschijningsdatum. *Fuzzy search*, *wildcards*, *ranking*: Valid/Search biedt de noodzakelijke flexibiliteit.

Het systeem is zeer schaalbaar en doorzoekt de collecties supersnel. De onderliggende technologie wordt ook gebruikt in bijvoorbeeld de Wikipedia.



Configureer, indexeer en publiceer

Het inzetten van Valid/Search verloopt in drie stappen: configureer, indexeer en publiceer.

 In een **configuratie** wordt vastgesteld waar uw gegevens zich bevinden. Files op uw netwerkschijf, bestanden via FTP, HTML documenten op een website, een MySQL of Oracle database.... alle zijn benaderbaar. Voorts geeft u aan hoe de gegevens moeten worden uitgelezen. Bijvoorbeeld: hoe is de auteur en titel van documenten in uw bibliografische database opgeslagen? En met welke frequentie moeten de bronnen opnieuw worden verwerkt? Daarnaast wordt de wijze waarop de gegevens moeten worden gerepresenteerd op het web vastgelegd.



Tijdens de **indexering** worden de gegevens uitgelezen. Als nodig worden stopwoorden verwijderd, stamvormen bepaald, trefwoorden uit de tekst onttrokken, e.d. Ook kan het document zelf als geheel worden bewaard. Hierdoor kan later – op het web – bijvoorbeeld snel een compleet PDF document worden aangeboden, of de lopende tekst in HTML vorm worden getoond.



Bij **publicatie** wordt de index op basis van zoekopdrachten in de webbrowser bevroegd. Niet alleen de zoekresultaten worden "on the fly" bepaald en getoond, maar de eigenschappen van de gegevensbronnen worden ook weergegeven. Het document zelf kan ook worden opgehaald ("download de PDF versie").

Bovenstaande functies worden ondersteund door de inzet van taaltechnologie. Dit betreft het identificeren, normaliseren en groeperen van belangrijke kenmerken binnen ongestructureerde tekst, zoals eigennamen en datums. Daarnaast wordt de bevraging uitgebreid met semantisch gerelateerde termen (inzet van thesaurus) en taalkundige stamvormen.

Niet of matig gestructureerde of gecategoriseerde informatie wordt hierdoor verrijkt en bruikbaar: "*Garbage In Quality Out*".

Open standaarden en open source

Deze oplossing is gerealiseerd door gebruik te maken van open source producten, en open standaarden. Hierdoor bent u minder afhankelijk van de leverancier en ontwikkelaars.

Het hart van het systeem is Lucene 2, een *high performance* ontsluitingssysteem dat in **open source** beschikbaar is en waar een groot aantal toepassingen op is ontwikkeld. Daarnaast wordt gebruik gemaakt van robuuste open source bibliotheken voor het *schedulen* van de indexering, het verwerken van XML bestanden, het uitlezen van de bronnen, *et cetera*. Valid/Search is zelf ook een open source product.

De aanpak die gevolgd wordt in Valid/Search is gebaseerd op XML, de **open standaard** van het W3C. De configuratie van het systeem wordt vastgelegd in XML documenten: de wijze waarop brongegevens worden vertaald naar een indexeerbaar formaat, de inrichting van het indexeringsproces, en de webapplicatie zelf. Hier wordt met name gebruik gemaakt van XSLT stylesheets waarin de hele webapplicatie kan worden vastgelegd.

Over Armatiek

Armatiek levert diensten op het terrein van informatieanalyse en -ontwerp, en realisatie van XML gebaseerde oplossingen. Onze technologie is onder andere ingezet door het Instituut voor Sociale geschiedenis (search.iisg.nl), Octrooicentrum Nederland (www.octrooicentrum.nl), en Technische Universiteit Delft (www.tudelft.nl).

Contactinformatie:

Armatiek BV, Krelis Louwenstraat 1 B08, 1055KA Amsterdam.
info@armatiek.nl / www.armatiek.nl

Stappenplan

De beschreven software kan worden ingezet voor kleine tot grote projecten. Een klein project zou kunnen zijn: ontsluit onze klantendatabase via het intranet. Een groot project kan zijn: geef toegang tot al onze documentatie in een uniforme, geïntegreerde website. Hoe pakt Armatiek zo'n project aan?

Stap 1: we bepalen welke gegevens zich waar bevinden, hoe deze logisch en fysiek zijn opgebouwd, en wat van deze gegevens beschikbaar moet worden gesteld via het web. We stellen vast hoe vaak de bronnen moeten worden benaderd om de actualiteit te waarborgen. Deze inzichten worden vastgelegd in een analysedocument.

Stap 2: hierna bepalen we op welke manier de gegevens moeten worden aangeboden, bijvoorbeeld: welke start- en detailpagina's zijn er, welke ingangen moeten er zijn op de zoekpagina, en hoe moeten de zoekresultaten worden weergegeven? Deze wensen omtrent de webapplicatie worden vastgelegd in een functioneel, grafisch en interactieontwerp.

Stap 3: als laatste stap installeren en configureren we het systeem. Besluiten rondom de bronnen worden vastgelegd in het configuratiebestand voor de indexer, en de raadpleegomgeving krijgt de vorm van een aantal XSLT scripts.