Provincie Zeeland

ZEEUWS ARCHIEF

MDTO-metadata met AI

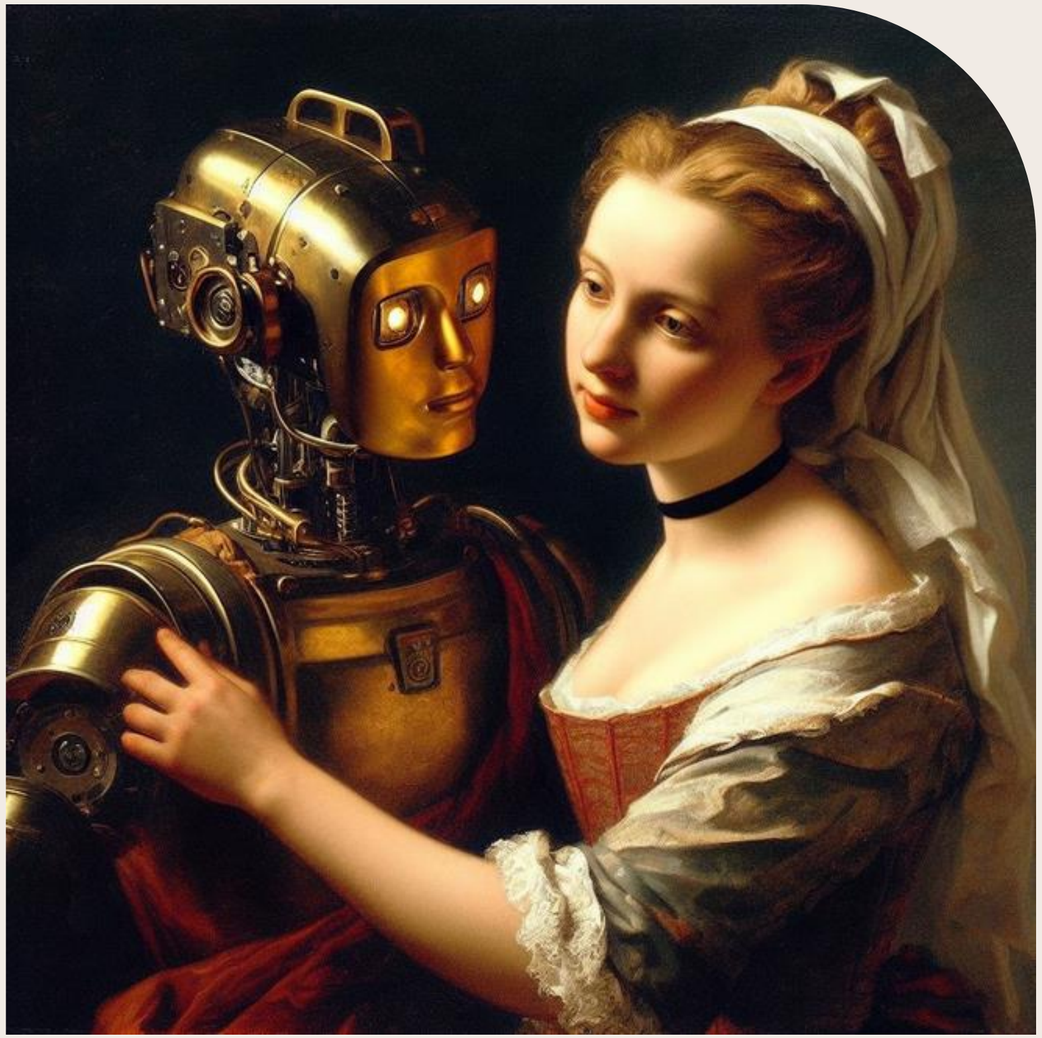
Van tekstuele informatie naar gestructureerde metadata in MDTO-formaat

Muriël Valckx

17 juni 2024

Tijdens deze presentatie

- ArchiveTextMiner
- Demo
- Feedback en ideeën



Afbeelding: NDE, gegenereerd met Dall-E

ALLE MOGELIJKE SOORTEN INFORMATIE

Provincie Zeeland

ZEEUWS ARCHIEF

Deel

1

ArchiveTextMiner

ALLE MOGELIJKE SOORTEN INFORMATIE

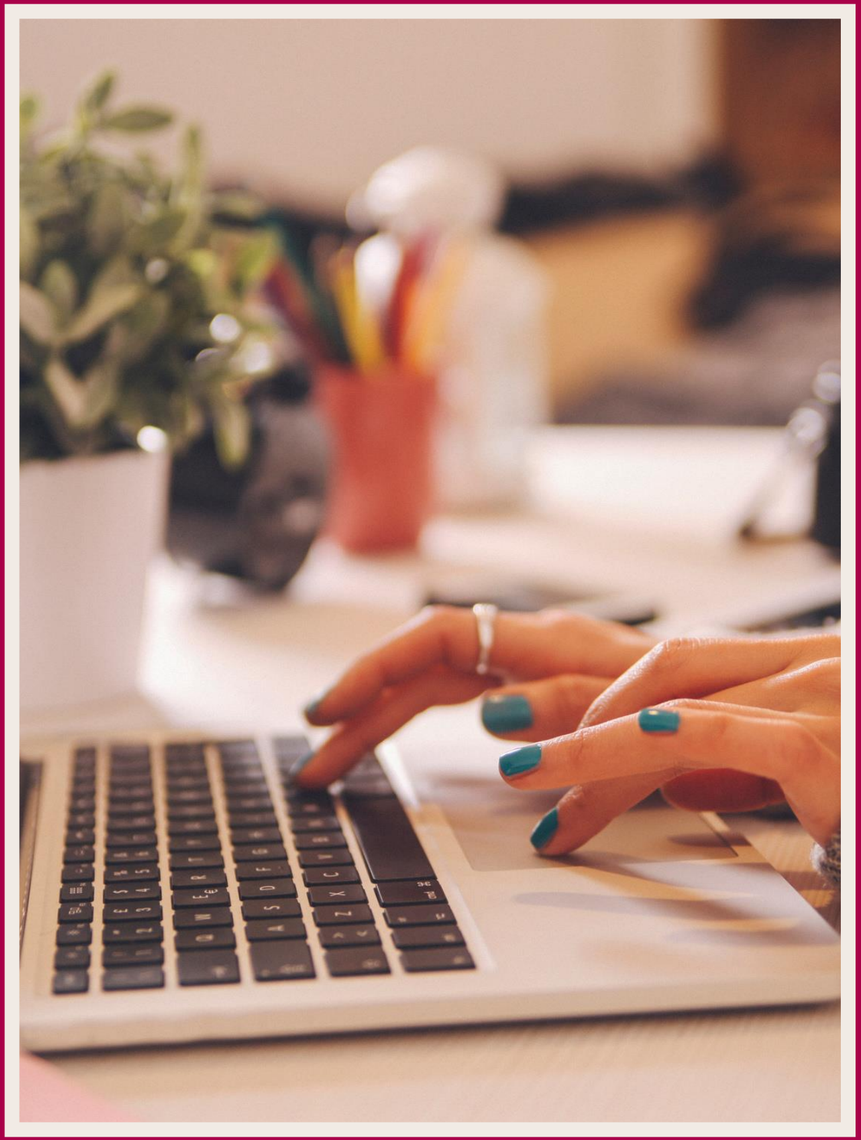
Provincie Zeeland

ZEEUWS ARCHIEF

ALLE MOGELIJKE
SOORTEN INFORMATIE

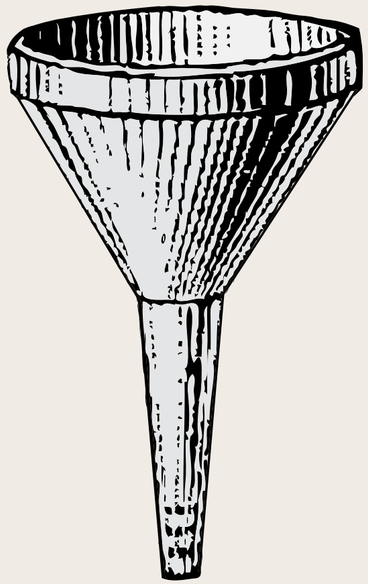
ChatGPT, ontbrekende metadata

“Kunnen we iets met text mining?”



ALLE MOGELIJKE SOORTEN INFORMATIE

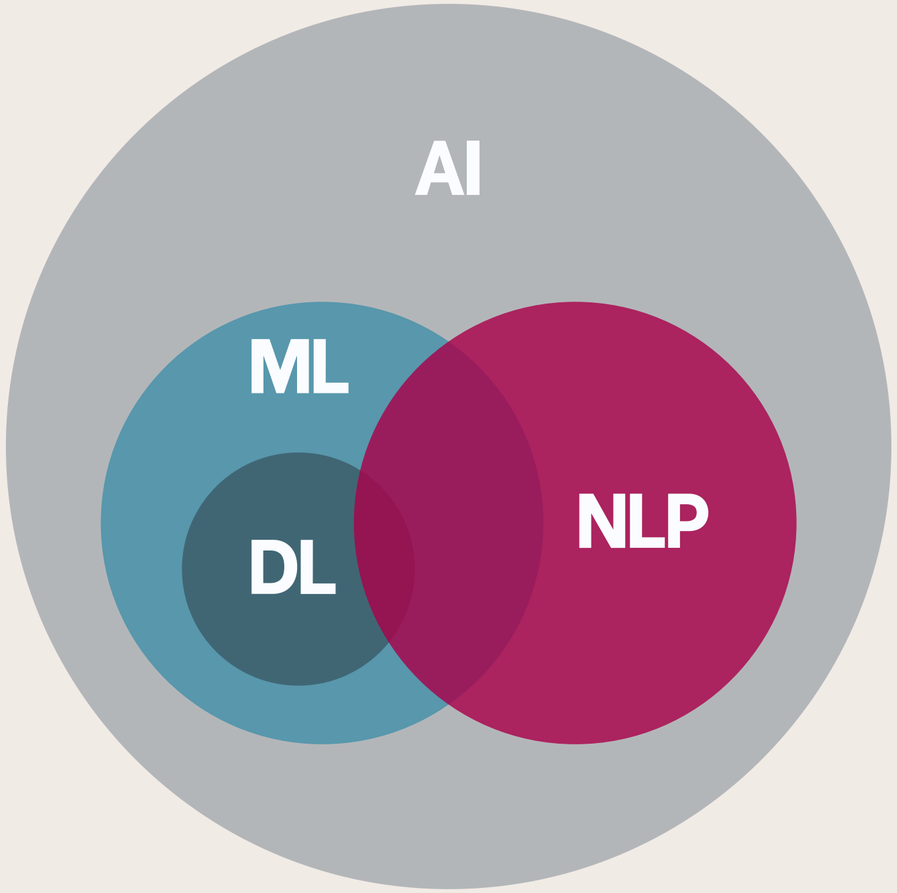
‘Inzichten verkrijgen uit grote hoeveelheden ongestructureerde tekst.’



ALLE MOGELIJKE SOORTEN INFORMATIE

Provincie Zeeland

ZEEUWS ARCHIEF



Artificial Intelligence (AI)
Computersystemen die taken kunnen uitvoeren die normaal gesproken menselijke intelligentie vereisen.

Machine Learning (ML)
Computersystemen die leren van gegevens en ervaringen.

Deep Learning (DL)
Computersystemen die het menselijk brein imiteren met complexe netwerken.

Natural Language Processing (NLP)
Computersystemen die menselijke taal begrijpen, interpreteren en genereren.

ALLE MOGELIJKE SOORTEN INFORMATIE
Provincie Zeeland
ZEEUWS ARCHIEF

Natural language processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity

Afbeelding: Huggingface.co

Hoe werkt text mining?

Opschonen

Inhoud analyseren

Nieuwe informatie

ALLE MOGELIJKE SOORTEN INFORMATIE

Provincie Zeeland

ZEEUWS ARCHIEF

ALLE MOGELIJKE
SOORTEN INFORMATIE

Tekst opschonen

De Prodentfabriek in Amersfoort werd in 1935 opgericht en produceerde tot 1971 tandpasta . Nu is HET een cultureel en creatief centrum met werkruimtes en /n historische industri&#235le elementen.

Tekst opschonen

De Prodentfabriek in Amersfoort werd in 1935 opgericht en produceerde tot 1971 tandpasta . Nu is HET een cultureel en creatief centrum met werkruimtes en /n historische industri&#235le elementen.

Tekst opschonen

Tokens:

["De", "Prodentfabriek", "in", "Amersfoort", "werd", "in", "1935", "opgericht", "en", "produceerde", "tot", "1971", "tandpasta", ".", "Nu", "is", "het", "een", "cultureel", "en", "creatief", "centrum", "met", "werkruimtes", "en", "historische", "industriële", "elementen", "."]

Tekst opschonen

Lemmatiseren:

['de', 'prodentfabriek', 'in', 'amersfoort', 'worden', 'in', '1935', 'oprichten', 'en', 'produceren', 'tot', '1971', 'tandpasta', '.', 'nu', 'zijn', 'het', 'een', 'cultureel', 'en', 'creatief', 'centrum', 'met', 'werkruimte', 'en', 'historisch', 'industrieel', 'element', '.']

Stemming:

['de', 'prodentfabriek', 'in', 'amersfoort', 'word', 'in', '1935', 'opricht', 'en', 'produceer', 'tot', '1971', 'tandpasta', '.', 'nu', 'zijn', 'het', 'een', 'cultureel', 'en', 'creatief', 'centrum', 'met', 'werkruimte', 'en', 'historisch', 'industrieel', 'element']

TF-IDF

Term Frequency-Inverse Document Frequency

Relevantie van woorden in een document, ten opzichte van een verzameling documenten.

1971	2.772.545
amersfoort	2.772.545
Prodentfabriek	2.772.545
1935	2.772.545
tandpasta	2.312.538
opgericht	1.846.441
produceerde	1.846.441
de	1.154.402
in	577.201
het	577.201
werd	577.201
tot	577.201
creatief	577.201
cultureel	577.201
centrum	577.201
werkruimtes	577.201
historische	577.201
industriële	577.201
elementen	577.201

ALLE MOGELIJKE SOORTEN INFORMATIE

Inhoud analyseren

```
root(werd)
    ├─── De (det)
    │       └─── Prodentfabriek (nsubj)
    ├─── in (case)
    │       └─── Amersfoort (obl)
    ├─── opgericht (advmod)
    │       └─── in (case)
    │               └─── 1935 (obl:tijd)
    ├─── produceerde (conj)
    │       ├─── en (cc)
    │       └─── tot (case)
    │               └─── 1971 (obl:tijd)
    ├─── tandpasta (obj)
    │       └─── 1971 (obl:tijd)
    ├─── Nu (advmod)
    │       └─── is (advcl)
    │               ├─── het (nsubj)
    │               └─── centrum (cop)
    ├─── een (det)
    │       └─── cultureel (amod)
    ├─── creatief (amod)
    ├─── centrum (nsubj)
    │       ├─── met (case)
    │       ├─── werkruimtes (conj)
    │       │       ├─── en (cc)
    │       │       └─── historische (amod)
    │       └─── elementen (conj)
    └─── . (punct)
```

Inhoud analyseren

root(werd): Het hoofdwerkwoord "werd" is de hoofd van de zin.
De (det): De bepaling "De" is afhankelijk van "Prodentfabriek".
Prodentfabriek (nsubj): Onderwerp van de zin.
in (case): Het voorzetsel "in" is afhankelijk van "Amersfoort".
Amersfoort (obl): Bijwoordelijke bepaling van plaats van "werd".
opgericht (advmod): Het voltooid deelwoord "opgericht" is een bijwoordelijke bepaling van "werd".
in (case): Het voorzetsel "in" is afhankelijk van "1935".
1935 (obl): Temporele bijwoordelijke bepaling van "opgericht".
produceerde (conj): Het werkwoord "produceerde" is verbonden met "werd".
en (cc): Voegwoord "en" verbindt "produceerde" en "tot".
tot (case): Het voorzetsel "tot" is afhankelijk van "1971".
1971 (obl): Temporele bijwoordelijke bepaling van "produceerde".
tandpasta (obj): Het lijdend voorwerp van "produceerde".
Nu (advmod): Het bijwoord "Nu" is een bijwoordelijke bepaling van tijd bij "is".
is (advcl): Een bijzin die het werkwoord "is" bevat.
het (nsubj): Het voornaamwoord "het" is het onderwerp van "is".
centrum (cop): Het koppelwerkwoord "centrum" is het gezegde van "is".
een (det): De bepaling "een" is afhankelijk van "cultureel".
cultureel (amod): Bijvoeglijk naamwoordelijk bepaling van "centrum".
creatief (amod): Bijvoeglijk naamwoordelijk bepaling van "centrum".
centrum (nsubj): Het onderwerp van "is".
met (case): Het voorzetsel "met" is afhankelijk van "centrum".
werkruimtes (conj): Geconjugeerd met "elementen".
en (cc): Voegwoord "en" verbindt "werkruimtes" en "historische".
historische (amod): Bijvoeglijk naamwoordelijk bepaling van "werkruimtes".
elementen (conj): Geconjugeerd met "werkruimtes".
. (punct): Punt aan het einde van de zin.

ALLE MOGELIJKE SOORTEN INFORMATIE

Nieuwe informatie

Resultaat metadataverrijking	
Originele bestandsnaam	20032474_S_20032474_4_A_43578_tds.pdf
Titel	December mensenrechtenvlag 10 mensenrechten
Samenvatting	De provincie Zeeland heeft een aanvraag ingediend bij het ministerie van Sociale Zaken en Werkgelegenheid om mee te doen aan de landelijke lancering van de Mensenrechtenvlag.
Steekwoorden	december, mensenrechtenvlag, 10, mensenrechten, provincie, nederland, lancering, vlag, hijsen, mee
Bestandsextensie	PDF document, version 1.4, 4 pages
Taal bestand	nl

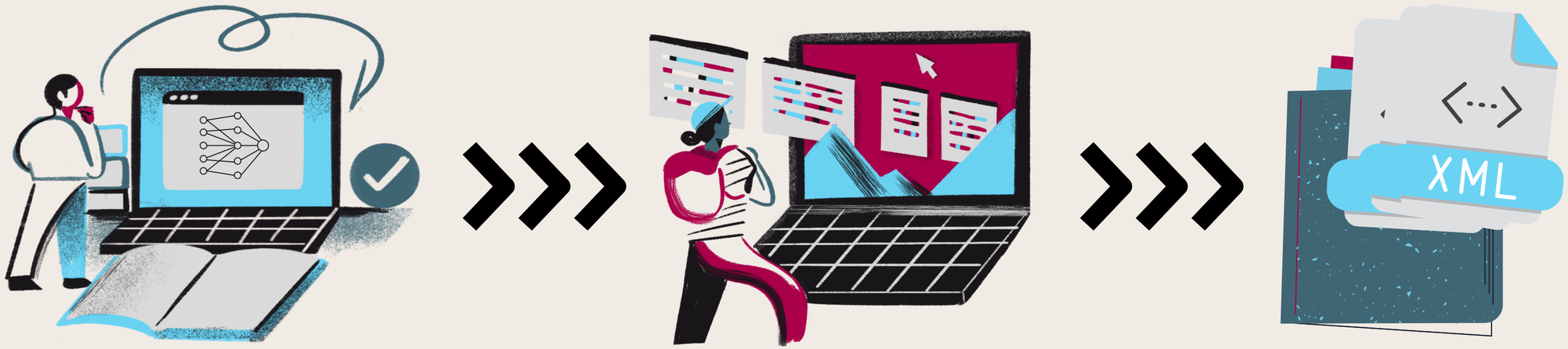
Resultaat metadataverrijking	
Originele bestandsnaam	GSnota ICL-IP_S_20032653_9_A_26978_tds.pdf
Titel	Iclip pva dcmr gs
Samenvatting	De provincie Zeeland (DCP) en de provincie Zuid-Holland (VVS) hebben in een vergadering van het ministerie van Veiligheid en Justitie een overkoepelend Plan van aanpak (PvA) van ICL-IP goedgekeurd.
Steekwoorden	iclip, pva, dcmr, gs, overtredingen, plan, bedrijf, situatie, toezicht, instanties
Bestandsextensie	PDF document, version 1.4, 4 pages
Taal bestand	nl

Resultaat metadataverrijking	
Originele bestandsnaam	20031924_S_20031924_1_A_50192_tds.pdf
Titel	Zeeland refinery mwe support
Samenvatting	Het ministerie van Economische Zaken en Waterstaat heeft de provincie Zeeland gevraagd om een 'letter of support' van de Provincie Zeeland voor het bedrijf Zeeland Refinery.
Steekwoorden	zeeland, refinery, mwe, support, letter, gs, waterstof, ', ', consequenties
Bestandsextensie	PDF document, version 1.4, 3 pages
Taal bestand	nl

Resultaat metadataverrijking	
Originele bestandsnaam	20032018_S_20032018_3_A_74348_tds.pdf
Titel	2020 begroting wijziging najaarsnota
Samenvatting	De effecten die betrekking hebben op het begrotingsjaar 2020 en 2021 zijn in het Statenvoorstel 10e wijziging begroting 2020, 2e wijziging begroting 2021 en 1e wijziging begroting 2021 vastgesteld.
Steekwoorden	2020, begroting, wijziging, najaarsnota, 2021, statenvoorstel, gs, zeeland, provincie, consequenties
Bestandsextensie	PDF document, version 1.4, 3 pages
Taal bestand	nl

Wat kan de tool?

- Géén ChatGPT, wel open source modellen
- Op veilige wijze informatie uit PDF-bestanden genereren



ALLE MOGELIJKE SOORTEN INFORMATIE

Provincie Zeeland

ZEEUWS ARCHIEF

Voorbeeld

```
python ArchiveTextMiner


    _          _    _        ____        _   __  __ _
   / \   _ __ ___| |__ (_)_   ___|_   _|___  _| |_|  \/  (_)_ __   ___ _ __
  / _ \ | '__/ __| '_ \| \ \ / / _ \| |/ _ \ \/ / __| |\/| | | '_ \ / _ \ '__|
 / ___ \| | | (__| | | | |\ V /  __/| |  __/>  <| |_| |  | | | | | |  __/ |
/_/   \_\_|  \___|_| |_|_| \_/ \___||_|\___/_/\_/\__|_|  |_|_|_| |_|\___|_|



Welkom bij ArchiveTextMiner.

ArchiveTextMiner starten...

Let op: De tekst in het veld 'omschrijving' is gegenereerd door kunstmatige intelligentie (AI). De betr

Bezig met uitpakken van input\bestanden.zip naar input\bestanden...

Uitpakken afgerond.

PDF-bestanden uitpakken in input\bestanden...

Bestanden ['Bestand-A.pdf', 'Bestand-B.pdf', 'Bestand-C.pdf'] gevonden.
Bezig met creëren metadata bestanden...

Bestand-A.pdf.xml succesvol gecreëerd.
Bestand-B.pdf.xml succesvol gecreëerd.
Bestand-C.pdf.xml succesvol gecreëerd.

Metadatabestanden opgeslagen in inputmap.

ArchiveTextMiner afgerond.
```

ALLE MOGELIJKE SOORTEN INFORMATIE

Provincie Zeeland

ZEEUWS ARCHIEF

Resultaat in MDTO-formaat


```xml
<MDTO xmlns="https://www.nationaalarchief.nl/mdto" xmlns_xsi="http://www.w3.org/2001/XMLSchema-instance" xsi_schemaLocation="https://www.nationaalarchief.nl/mdto https://www.nationaalarchief.nl/mdto/MDTO-XML1.0.1.xsd">
  <bestand>
    <identificatie>
      <identificatieKenmerk />
      <identificatieBron />
    </identificatie>
    <naam>IB19_359da5ea.pdf</naam>
    <omvang>154641</omvang>
    <omschrijving>De Zeeuwse Natuur- en biodiversiteitorganisatie Zeeuwse Natuur- en biodiversiteitorganisaties (Z.M.A.) en het Zeeuwse Natuur- en biodiversiteitinstituut (Z.M.A.) hebben een commissie geleid die adviseert</omschrijving>
    <taal>nl</taal>
    <kvkNummer />
    <bsnNummer />
    <bestandsformaat>
      <begripLabel>PDF document, version 1.4</begripLabel>
      <begripCode />
      <begripBegrippenlijst>
        <verwijzingNaam>nota, natuurbeleving, ps, commissie, in, bedenkingen, de, gs, maart, natuur</verwijzingNaam>
      </begripBegrippenlijst>
    </bestandsformaat>
    <checksum>
      <checksumAlgoritme>
        <begripLabel />
        <begripBegrippenlijst>
          <verwijzingNaam>Nota Natuurbeleving Ps</verwijzingNaam>
        </begripBegrippenlijst>
      </checksumAlgoritme>
      <checksumWaarde />
      <checksumDatum />
    </checksum>
    <URLBestand />
    <isRepresentatieVan>
      <verwijzingNaam />
      <verwijzingIdentificatie>
        <identificatieKenmerk />
        <identificatieBron />
      </verwijzingIdentificatie>
    </isRepresentatieVan>
  </bestand>
</MDTO>
```

ArchiveTextMiner

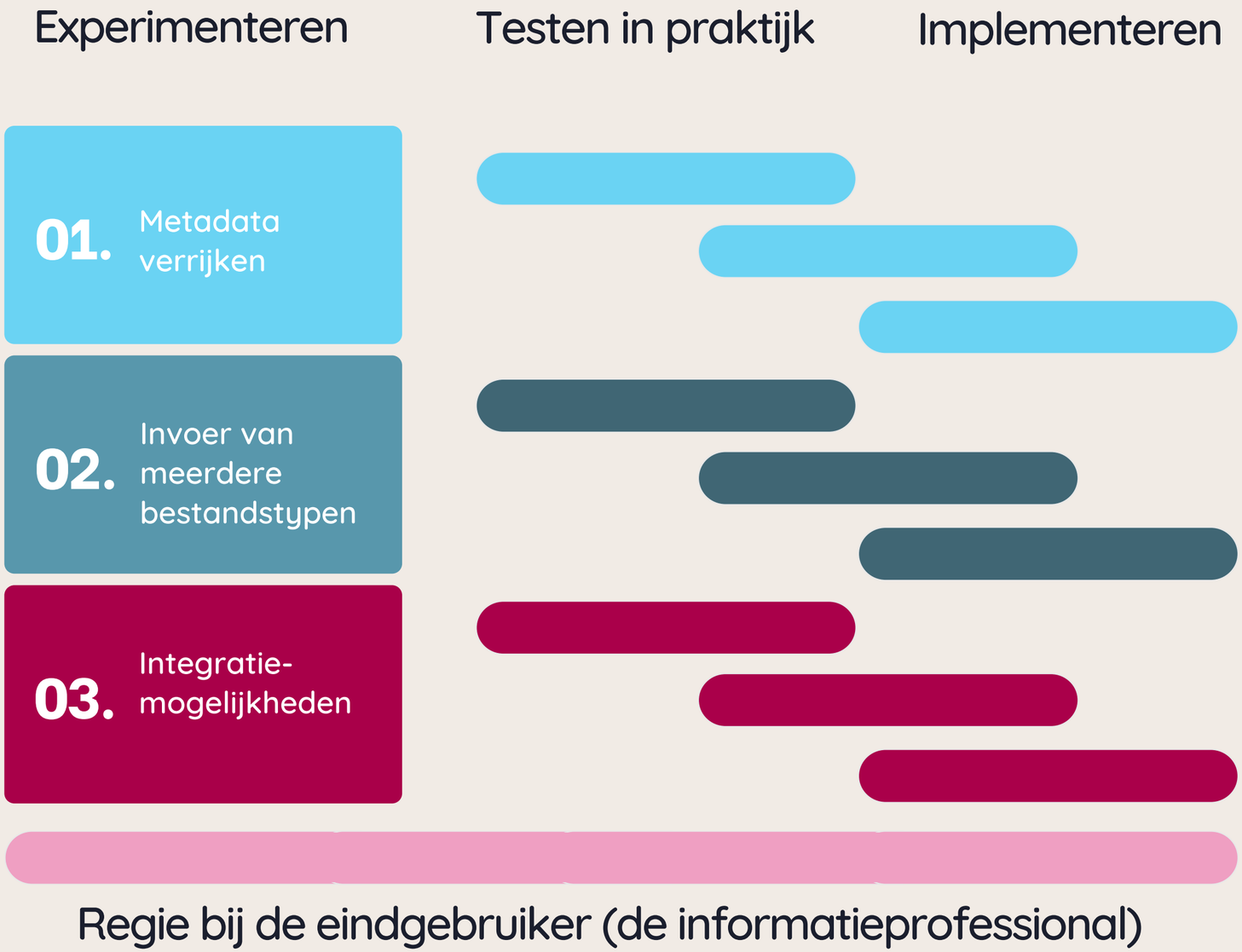
https://github.com/zeeuws-archief/ArchiveTextMiner

ALLE MOGELIJKE SOORTEN INFORMATIE

Provincie Zeeland

ZEEUWS ARCHIEF

Toekomst



ALLE MOGELIJKE SOORTEN INFORMATIE

Provincie Zeeland

ZEEUWS ARCHIEF

Deel

2

ArchiveTextMiner

Demo

ALLE MOGELIJKE SOORTEN INFORMATIE

Provincie Zeeland

ZEEUWS ARCHIEF

https://es-connector.acc.zeeland.nl/atm_demo

ALLE MOGELIJKE SOORTEN INFORMATIE

Provincie Zeeland

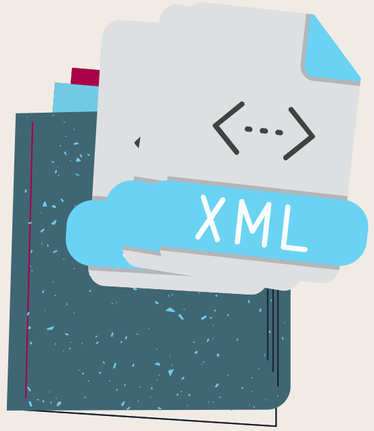
ZEEUWS ARCHIEF

Community

Invoe

Gebruikerservarin

Uitvoer



Mentimeter

Interesse gekregen?
my.valckx@zeeland.nl

Enquête voor feedback en mogelijkheid aanmelden voor community:

https://www.survio.com/survey/d/M2M8O7E8Y5D1V6J3A