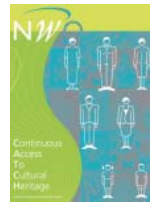
# Structuren in de chaos?

**Mogelijkheden voor de informatiewetenschap binnen het cultureel erfgoed**

**Paul Doorenbosch (KB)**



Continous

Access

To

Cultural

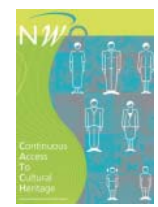
Heritage

Een NWO informaticaonderzoeksprogramma

voor en met het

Cultureel Erfgoed





Cultureel Erfgoed:

alle bronnen uit het verre en nabije verleden waarvan we vinden dat het iets interessants of leuks over ons verleden kan vertellen en die we de moeite van het bewaren waard vinden:

schilderijen, fossielen, landschappen, monumenten, poëzie, muziek, foto’s, borden, kleren, flessen, scherven, botten, speelgoed, vliegtuigen, kranten, ...


taal

# Het belang van het gedigitaliseerde cultureel erfgoed

Bronnen en
kennis over bronnen
samengebracht

Nieuw (alfa-)onderzoek

Het publieke belang en de
Verborgen schatten

Het is van ons en voor ons

Een groot deel van ‘het leven’ gaat zich in een
digitale omgeving afspelen

Wat is er mis met het huidige digitale aanbod?

Waarom niet gewoon alles maar op een giga harde schijf, html-voorkant en hup, internet op!

- Arme of geen metadata
- Gebrekkig vinden in teksten
- Veel resultaten, maar eigenlijk niet wat je zoekt
- Gebrekkige navigatie
- Mensonvriendelijke interfaces
- Weinig combinatiemogelijkheden

- Veelal collectiegebonden
- Gebrekkige context
- Slechte inzetbaarheid van bestaande kennis
- Massaliteit
- Wat kan ik hierin vinden?
- Ontsluiting niet-tekstmateriaal
- Weinig geld en weinig mensen

Wat hebben we aan informaticaonderzoekers?

Wat moeten we met een proefschrift over vier/vijf jaar?

| HUIDIGE SITUATIE: | VERWACHTING: |
| --- | --- |
| Pragmatisme | Vanuit een ander hoek bekijken |
| Remmende voorsprong | Andere methoden en technieken |
| Begrensde horizon | Nieuw(st)e informatica-kennis |
| Vanuit een vakinhoudelijke of erfgoedperspectief | Samenwerking / Wederzijdse beinvloeding |
| | Vernieuwende kennis EN praktische toepassingen |


MultimediaN

MultimediaN

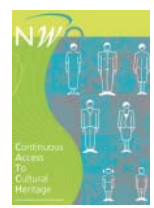
**Multimedia Netherlands**

involves the knowledge creation and transfer on handling of video, pictures, audio and language in ICT

deelproject: application pilot in the field of e-culture

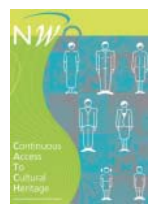
Despite the amounts of public funding devoted to both cultural heritage and ICT infrastructure, online access to even the most important aspects of our past is still limited and highly fragmented. The objective of this project is the development of a set of e-culture demonstrators providing multimedia access to distributed collections of cultural heritage objects. The demonstrators are intended to show various levels of syntactic and semantic interoperability between collections and various types of personalized and context--dependent presentation generation.



# deelproject: semantic access

The semantic multimedia access project concentrates on the development of generic technology that satisfies multimedia information at a semantic level. Any search system's comprehension of a user's information need is necessarily incomplete, as this would require understanding completely the user's goals as well as the user’s perception of retrieved objects. The project investigates how in spite of this uncertainty effective search strategies can be offered, exploiting the following search parameters: the collection (domain knowledge, background knowledge, language, format, etc.), the user (including use scenario, interaction type, history, preference) and the system (security aspects, performance). Users often know things about the multimedia objects in a collection being searched; yet, it remains a challenge how to exploit and adapt to such background knowledge during search. In this project we develop a search engine generator based on probabilistic retrieval models.
The project will focus on technology and tools applicable in the domain of media and e-culture and ambient settings.



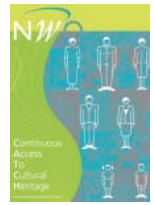
# CATCH

## DRIE HOOFDLIJNEN VAN ONDERZOEK

a. Semantische interoperabiliteit via metadata.

b. Kennisverrijking met behulp van automatische analyse.

c. Personalisatie in presentatie.




# Zes kernprojecten

**STICH**: metadata-interoperabiliteit door semantische verbindingen

**CHOICE**:semi-automatische annotatie met audio-visuele informatie

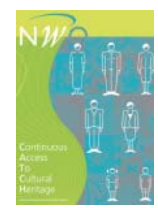
**RICH**: automatische herkenning en classificatie van archeologische voorwerpen

**SCRATCH**: zoeken in handgeschreven archieven

**MITCH**: logboeken omzetten in verrijkte databases

**CHIP**: gepersonaliseerde rondleidingen in virueel museum



# **STICH**: metadata-interoperabiliteit door semantische verbindingen

Belangrijkste uitdagingen:

• Hoe kunnen we de handmatig aangebrachte structuren in ontologieën ten volle benutten voor automatische verwerking?

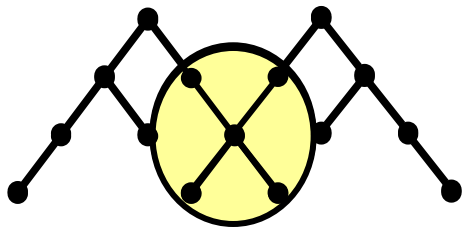
• Hoe kunnen we de kennis van specialistische ontologieën breder inzetbaar maken?

• Hoe kunnen we ontologieën (breed of smal) verknopen ten behoeve van meer mogelijkheden voor retrieval?

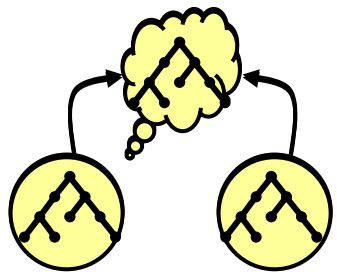
• Hoe kunnen we verknoopte thesauri inzetten bij semantische analyse van volledige teksten?

**STICH**: metadata-interoperabiliteit door semantische verbindingen

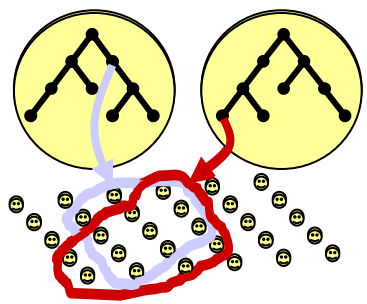
- ‘Shared’ vocabularies



- ‘Upper-level’ ontology



- ‘Shared Instances’



**STICH**: metadata-interoperabiliteit door semantische verbindingen

- Onderzoeksuitdaging: ‘ontology mapping’
- Metadata interoperabiliteit via semantische links
- Onderszoeksvragen:
  - Onderscheiden soorten semantische links?
  - Hoe deze te identificeren (handmatig/semi-automatisch?)
  - Hoe te gebruiken voor toegang tot heterogene collecties?)

**STICH**: metadata-interoperabiliteit door semantische verbindingen

- Eindgebruiker: Verfijnde Vindhulpmiddelen
- Efficiënte Metadata Onderhoudstools
- Koppeling diverse Collecties and Vocabulaires

**RICH**: automatische herkenning en classificatie van archeologische voorwerpen

## Four challenges

1. How can we safeguard the existing knowledge base?
2. How can we guarantee fast and easy access for all?
3. How can we guarantee the incorporation of new knowledge in a sustainable way?
4. How can we enrich the existing and forthcoming knowledge by new techniques?

**INTENSIFIED and DIVERSIFIED COMMUNICATION and KNOWLEDGE EXCHANGE**

**RICH**: automatische herkenning en classificatie van archeologische voorwerpen

## Aims of RICH

- Increasing the efficacy and efficiency of digital access to archaeological core knowledge.
- Reinforcing the infrastructure on archaeological core knowledge.
- Improving the quality of material studies in Dutch archaeological heritage management and archaeological research in Europe, including the formulation of new research area's.

**RICH**: automatische herkenning en classificatie van archeologische voorwerpen

Main research question

- How can artificial intelligence support the automatic visual analysis of archaeological objects?

**RICH**: automatische herkenning en classificatie van archeologische voorwerpen

## Approach and methodology

- An empirical approach based on image-(pre)processing and machine-learning techniques

- The scientific methodology (four phases):
  1. data collection,
  2. data pre-processing,
  3. training, and
  4. evaluation.

## Toegankelijkheid van handgeschreven archieven

- digitaliseren (scannen van het beeld) verhoogt de toegankelijkheid van teksten niet!
- handmatige menselijke transcriptie op woordniveau (naar “Word” bestand) is te duur
- automatische herkenning van handschrift (OCR) is maar in zeer beperkte mate mogelijk:

  Verbonden schrift is nauwelijks machinaal te interpreteren

Echter…

- ook al werkt machinale transcriptie niet:
- zoekmethoden in tekst- en beeldmateriaal zijn sterk verbeterd (“Information Retrieval”)
- rekenkracht van computers is toegenomen
- patroonherkenning (berekenen van schriftkenmerken) wordt snel beter

**SCRATCH**: zoeken in handgeschreven archieven

Googelen in handgeschreven archieven?
drie mogelijke scenario's

- a) steekwoorden: de gebruiker omlijnt met de muis een handgeschreven woord, tikt de bijbehorende tekst in en de computer gaat op zoek
- b) gebruiker klikt op handgeschreven woordbeelden: de computer gaat op zoek
- c) gebruiker omlijnt een paragraaf: de computer gaat op zoek naar paragrafen met een soortgelijke inhoud

NB: In scenario a) leert de computer van elke zoekopdracht !
Het zoeken zal steeds beter gaan.

Techniek onder de motorkap

- Zelf-organiserende leermethoden toepassen op vormkenmerken van letters en woordfragmenten
- Toepassen van kennis over de document-structuur van het Kabinet van de Koning
- Toepassen van kennis over Nederlandse taal
- Gebruikmaken van “bag of words” methoden uit “Information Retrieval” ➔ “bag of written shapes”

# **CHOICE**:semi-automatische annotatie met audio-visuele informatie

- Semantisch annoteren van met name video
- Interfaceontwikkeling:
- Gebruik van ontologieen
- Gebruik van NLP-technieken voor semantische categorisering
- Ondersteuning van zoekproces met resultaten hiervan

# **MITCH**: logboeken omzetten in verrijkte databases

- Schonen data in database

- Semantische relaties leggen met begrippen binnen database en interne begrippenbestanden

- Relateren van termen, phrases en domein-specifieke velden naar achtergrond teksten (intern en extern)

# **CHIP**: gepersonaliseerde rondleidingen in virueel museum

- Presentatie – navigatie - personalisatie
- Heterogene bestanden en combinaties van bestanden
- Verschillende niveau’s presenteren vanuit zelfde bronnen
- Delen van gebruikerskenmerken met verschillende instanties

## CATCH: voor wie in het erfgoed?

Cultureel en historisch geïnteresseerden

Onderzoekers

maar in de eerste plaats gericht op:

Beheerders van erfgoedcollecties (als de intermediairs)

## Waarom?

Interactie tussen publiek en verleden

MultimediaN: vanuit de techniek

CATCH: vanuit het Cultureel Erfgoed

spanning (voor het CE) in
beide programma's:

informatica-onderzoek

versus

software-engineering

‘PROBLEMEN’ in het erfgoed,

waaraan de informatica mogelijk een bijdrage zou kunnen leveren

# ‘PROBLEMEN’ in het erfgoed

Essentiele kennis uit hoofden halen en machine-interpretabel vastleggen

Democratisering van het erfgoed

Metadatatoekenning

Menselijke expertise is niet efficient

Massaliteit van het erfgoed – schaalbaarheid – managen

Voorbij de menselijke maat

Robuustheid en accuraatheid gaan slecht samen (Antal)

Navigatie in grote verzamelingen

# ‘PROBLEMEN’ in het erfgoed

Domeinspecificiteit vs brede belangstelling

Retorische taal, impliciete taal, ambigue taal

Taalregisters

OCR / post-ocr

Kennisextractie uit teksten

Interfaces / pda’s /etc

Personalisatie – localisatie, GIS, RFID

Interactief zoeken – vraaggeleid zoeken  - dialoog

# ‘PROBLEMEN’ in het erfgoed

Betrouwbaarheid van leverancier

van data en dataverzamelingen

Veranderende autoriteit

DRM – beveiliging (hergebruik/stelen)

Verbetering bedrijfsprocessen / efficientie

Interoperabiliteit data en metadata:

standaardisering versus ontwikkeling =?

stilstand versus vooruitgang

# ‘PROBLEMEN’ in het erfgoed

Semantisch web: toekomst, betekenis.

taxonomieen

ontologieen

thesauri

(gecontroleerde) trefwoordenlijsten

intellect versus machine

Beeldherkenning

Beeldgrammatica

Visueel zoeken (beeldzoeken)

Bewaren, hergebruiken, ontsluiten van born digital (kunst, archief, software, …..)

Digitale duurzaamheid

# ‘PROBLEMEN’ in het erfgoed

Bedreiging van informatica voor erfgoed

Biedt informatica oplossingen / nieuwe kansen?

Achter de feiten / ontwikkelingen in de wereld aanlopen

Kennis vergaren / buitenwereld volgen

# ordening in de chaos

De wereld waarin wij leven is ‘ogenschijnlijk’ *wanordelijk en gefragmenteerd?*

CE instellingen willen daar graag *orde, structuur en samenhang* in brengen?

Vraagt het *publiek* ook om die orde?

Is de orde van de *beheerder* dezelfde orde als die van het publiek?

Is die orde hetzelfde in elke *omstandigheid, tijd en plaats*?

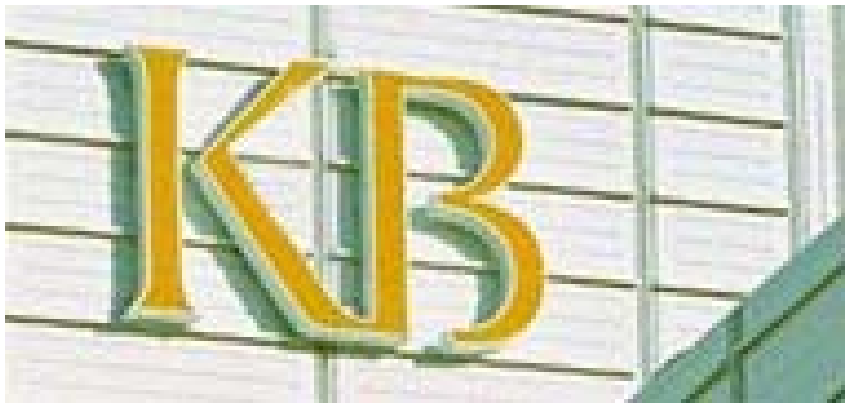
Het publiek kan omgaan met de wanorde van de ‘echte’ wereld,
waarom dan niet met die van het *digitale erfgoed?*

Kunnen wij op de pc de kennisordening van een persoon *representeren?*

*HIERIN KAN DE INFORMATICA EEN BELANGRIJKE ROL SPELEN*

Met elkaar praten
Naar elkaar luisteren
Elkaars taal willen spreken
Met elkaar werken
Van elkaar willen leren
Kennis aanreiken
Kennis delen
Over de muren kijken
Krachten bundelen
Samenwerken

**Betere interactie tussen publiek en digitaal erfgoed**



paul.doorenbosch@kb.nl