Oudezijds Achterburgwal 185
1012 DK Amsterdam
Postbus 10855
1001 EW Amsterdam
Tel: +31 20 224 6800
info@huygens.knaw.nl
www.huygens.knaw.nl

MEMO

Datanotitie Huygens Instituut voor Nederlandse Geschiedenis

Datum: 14 februari 2017
Datanotitie Huygens ING versie 2

Contactpersoon: Sebastiaan Derks
Telefoonnummer: 020 224 68 52
sebastiaan.derks@huygens.knaw.nl

Inleiding
De rol en betekenis van onderzoeksdata binnen de geesteswetenschappen veranderen in hoog tempo. Door digitalisering van historische documenten, literaire werken, gestructureerde gegevens, en visuele objecten komen enorme hoeveelheden data beschikbaar. Dankzij digitale technologieën hebben onderzoekers bovendien ongekende mogelijkheden om deze gegevens te analyseren en nieuwe, grotere, discipline-overstijgende verbanden te leggen. Maar om daadwerkelijk te kunnen profiteren van de overweldigende vloed aan onderzoeksdata, zijn inhoudelijke bewerking en onderhoud noodzakelijk. Dit geldt in het bijzonder voor de geesteswetenschappen, waar onderzoeksdata worden gekenmerkt door een enorme diversiteit en heterogeniteit: gegevens verschillen vaak per locatie, periode en sociale context, en vertonen bovendien vaak grote variaties in vocabulaire, vorm en volledigheid. Alleen wanneer deze data met elkaar worden verbonden en de idiosyncratische ordeningen worden omgezet naar gemeenschappelijke structuren, kunnen ze goed worden ingezet voor nieuw onderzoek. Hierdoor is datamanagement een basisvoorwaarde voor goed geesteswetenschappelijk onderzoek geworden. Voor zover onderzoekers dit niet zelf al erkenden, worden zij door de nationale en Europese onderzoeksfinanciers steeds vaker verplicht om hun data adequaat te beheren en de toegankelijkheid ervan te waarborgen. Kortom, het delen van data speelt een steeds grotere rol in de onderzoekspraktijk van de geesteswetenschappen.

Tegen deze achtergrond heeft het Huygens ING, dat een sterke traditie kent van data-intensief historisch en letterkundig onderzoek, begin 2014 besloten om de activiteiten op het terrein van databeheer en onderzoeksassistentie te clusteren in een afdeling Digitaal Databeheer. Haar primaire verantwoordelijkheid is de ontwikkeling en duurzame beschikbaarstelling van digitale databestanden en resources voor geesteswetenschappelijk onderzoek. De afdeling beoogt het gebruik van data door onderzoekers te vergemakkelijken en nieuw onderzoek te stimuleren. Dit doen wij door het verzamelen en bundelen van digitale onderzoeksgegevens binnen innovatieve, gedeelde data-infrastructuren. Het doel van deze infrastructuren is onderzoekers in staat te stellen geesteswetenschappelijke data, vooral op het gebied van de Nederlandse geschiedenis, op een toegankelijke wijze te delen, te verifiëren en te analyseren.

Tijdens het werken met deze infrastructuren zijn er steeds methodologische en organisatorische kwesties die buiten de scope van afzonderlijke onderzoeksprojecten vallen. Hoe valideren we datasets zo dat ze ook inzetbaar blijven voor andere onderzoeksvragen? Hoe blijft de herkomst en kwaliteit van onderzoeksdata, zelfs na intensieve bewerkingen door verschillende partijen, inzichtelijk en controleerbaar? Wat zijn de rechten van en plichten voor onderzoekers bij het delen van hun onderzoeksgegevens? Welke formats en ontologieën zijn het meest geschikt voor bepaalde typen onderzoek? Hoe kunnen we het proces van informatie-extractie uit historische en

letterkundige bronbestanden verbeteren? Hoe kunnen we vraaggestuurd werken zonder aan uniformiteit en coherentie in te boeten?

De complexiteit van deze vraagstukken vereist de inzet van een specifiek type professionals – data-officers en -managers – met gedegen domeinkennis en voldoende technisch inzicht om alle aspecten goed af te wegen en het proces te managen. Bij het Huygens ING werken we steeds in kleine projectteams waarin verschillende expertises (onderzoek, datamanagement en IT) worden samengebracht. De data-officers zijn daarbij verantwoordelijk voor het inhoudelijk functioneren van de data-infrastructuren. In veel opzichten – zowel inhoudelijk als organisatorisch – kunnen data-officers daarom worden getypeerd als de beheerders van geesteswetenschappelijke laboratoria.

Datanotitie
Deze datanotitie beschrijft hoe het Huygens ING invulling geeft aan de KNAW-dataprincipes en -beleid[1], gegeven de eigen onderzoeksdoelstellingen. Het gaat hierbij alleen over data die verzameld of gecreëerd zijn in het kader van Huygens ING-projecten en over data van externe onderzoeksprojecten die zijn opgeslagen in de Huygens ING-datainfrastructuren. De datanotitie zal jaarlijks worden geëvalueerd en indien nodig herzien opdat zij goed aansluit op de best practices in datamanagement en op de ontwikkelingen in de geesteswetenschappen.

Dataprincipes
Datamanagement bij het Huygens ING vindt plaats volgens de voorwaarden van de (onderzoeks)financiers en volgens de onderzoekpraktijken en -codes van de geesteswetenschappen. Vanzelfsprekend nemen wij bij ons werk de relevante wet- en regelgeving (bijv. de Auteurswet en de Wet Bescherming Persoonsgegegevens) in acht. Het Huygens ING volgt voor zijn dataprincipes de KNAW-uitgangspunten (die deels overeenkomen met de FAIR-dataprincipes: Findable, Accessible, Interoperable, Reusable), maar brengt daarbij eigen (domeinspecifieke) accenten aan[2]:

1. Toegankelijkheid
Geesteswetenschappelijke data die zijn geproduceerd in het kader van Huygens ING-projecten of in de infrastructuren van Huygens ING zijn opgeslagen, beschouwen wij als publiek goed en zijn vrij beschikbaar voor anderen om te gebruiken. Het Huygens ING staat dus voor onderzoeksdata die open-access toegankelijk zijn. Uitzonderingen op dit principe zijn mogelijk in verband met redenen rond privacy, rechten van derden, staatsbelangen, etc.[3], en in de vorm van een tijdelijk embargo op de data voor onderzoekers (uiteraard alleen indien dit niet conflicteert met de voorwaarden van onderzoeksfinanciers en -instellingen).

2. Provenance
Datainfrastructuren in de geesteswetenschappen ontwikkelen zich steeds meer tot ‘levende ecosystemen’, waarin onderzoeksdata door verschillende partijen, onafhankelijk van elkaar, worden bewerkt en gebruikt. Dit heeft tot gevolg dat veel van de acties binnen de data-lifecycle[4] niet meer direct gerelateerd zijn aan afzonderlijke onderzoeksprojecten, maar juist door het delen van onderzoeksdata gemeenschappelijk zijn geworden. Het vaak gemaakte onderscheid tussen datamanagement ‘tijdens’ en ‘na’ het onderzoek wordt hierdoor geleidelijk ook moeilijker te hanteren. Belangrijker zijn de inhoudelijke en methodologische gevolgen van deze ontwikkeling: de nieuwe mogelijkheden om onderzoeksdata te aggregeren en te analyseren roepen ook nieuwe vragen op over de herkomst van informatie (provenance). Voor geesteswetenschappers is het

[1] https://intranet.knaw.nl/nl/ict/databeleid/160909-KNAW-dataprincipes%20en%20KNAW-databeleid.pdf
[2] Vanzelfsprekend bieden wij nadere toelichting wanneer we in uitzonderlijke gevallen toch moeten afwijken van deze principes.
[3] Zie de lijst van de Nederlandse overheid met redenen om een dataset gesloten te houden: https://data.overheid.nl/reden-gesloten.
[4] http://www.dcc.ac.uk/resources/curation-lifecycle-model

essentieel om zo exact mogelijk te weten welke bronnen, applicaties en methoden zijn toegepast om een bepaald onderzoeksresultaat te genereren. Dat is immers de enige weg om zicht te krijgen op de representativiteit, kwaliteit, en reproduceerbaarheid ervan. Het Huygens ING maakt zich daarom sterk voor het nauwgezet documenteren van dataprovenance in haar meest ruime betekenis: van de selectiecriteria van de oorspronkelijke data tot alle daaropvolgende bewerkingen, van de complexe algoritmen waarmee automatische koppelingen worden aangebracht tot de verantwoording van de gemaakte keuzes bij informatie-extractie.[5] We streven dus naar transparantie bij alle datasets en acties, geven hiermee digitale hermeneutiek de ruimte zich te ontwikkelen en stellen nieuwe normen voor wetenschappelijke betrouwbaarheid.

3. Duurzaamheid
Het Huygens ING werkt voortdurend aan duurzame oplossingen voor de opslag van geesteswetenschappelijke data. Onderzoeksgegevens opslaan lijkt wellicht eenvoudig, maar is een zeer bewerkelijk en complex traject dat binnen goed doordachte infrastructuren moet worden uitgevoerd om duurzaamheid en stabiliteit te kunnen waarborgen. Data die relevant zijn voor geesteswetenschappelijk onderzoek worden bij het Huygens ING opgeslagen en beschikbaar gesteld voor de lange termijn via de eigen datainfrastructuren, die zijn gecertificeerd als trusted digital repositories. Deze infrastructuren (een voor gestructureerde data en een voor tekstuele data) draaien op eigen softwaresystemen: Timbuctoo en Alexandria. Beide zijn voorzien van een gebruiksvriendelijke interface en up- en downloadfunctionaliteit die de interactie met onderzoekers bevorderen. Timbuctoo en Alexandria maken gebruik van innovatieve Linked Open Data-technologieën en worden constant doorontwikkeld door het Huygens ING. De twee softwaresystemen zijn open-source en vrij beschikbaar onder GPL v3. Duurzame oplossingen in dit tijdperk van linked-data vragen echter ook om goede samenwerking en afstemming tussen verschillende partijen. Het Huygens ING besteedt daarom veel aandacht aan nationale samenwerkingsverbanden (bijv. KNAW Humanities Cluster, CLARIAH, Netwerk Digitaal Erfgoed, Landelijk Coördinatiepunt Research Data Management) en internationale initiatieven op het terrein van data-infrastructuren.

Datasoorten
In geesteswetenschappelijke onderzoeksdata zijn ruwweg drie soorten te onderscheiden: de digitale objecten (in het geval van het Huygens ING veelal document-images, OCR-bestanden, en XML-transcripties), gestructureerde gegevens over deze objecten (metadata, contextuele informatie, annotaties, provenance), en conceptuele data (classificaties, netwerkschema's, thesauri, vocabulaires, ontologieën, etc.). In principe zijn al deze datasoorten relevant voor hergebruik door het Huygens ING: nieuwe technologieën en aanvullende of verwante data kunnen immers weer tot nieuwe onderzoeksresultaten leiden.

Dataparagraaf en DMP's
Het Huygens ING ziet datamanagement als een integraal onderdeel van wetenschappelijk onderzoek. Daarom biedt de afdeling Digitaal Databeheer onderzoekers van het instituut en zijn partners ondersteuning en advies in alle fasen van het verzamelen en aggregeren van onderzoeksgegevens. Vanwege de grote diversiteit en heterogeniteit van geesteswetenschappelijke data zien we daarbij weinig in generieke oplossingen (bijv. templates voor de dataparagraaf of datamanagementplannen – DMP's[6]), maar kiezen we nadrukkelijk voor het bieden van maatwerk. De afdeling helpt onderzoekers bij de implementatie van landelijke en internationale datamanagementprotocollen: wij ondersteunen onderzoekers actief bij het opstellen van een dataparagraaf. Het toespitsen van

[5] Vgl. voor deze aanpak: Niels Ockeloen, Antske Fokkens, Serge ter Braake, Piek Vossen, Victor de Boer, Guus Schreiber and Susan Legêne (2013) BiographyNet: Managing Provenance at multiple levels and from different perspectives. In: *Proceedings of the Workshop on Linked Science* (LiSC) at ISWC 2013, Sydney, Australia, October 2013.
[6] Zie bijv. https://dmponline.dcc.ac.uk/

datamanagement op de karakteristieken van het onderzoek en de aansluiting op de infrastructuren van het instituut maakt de aanvraag sterker. Recente honoreringen lijken dit voordeel voor onderzoekers ook te bevestigen. Na aanvang van het onderzoek nemen vertegenwoordigers van de afdeling ook zitting in het projectteam. Deze data-officers zijn verantwoordelijk voor de kwaliteitsborging en inhoudelijke functionaliteit van de datainfrastructuren. Zij blijven ook gedurende de looptijd van het project nauw betrokken: zij stellen in overleg met het team het datamanagementplan op en houden dit up-to-date, geven advies over structurering van gegevens, vertalen complexe onderzoeksvragen naar concrete data-mappings en zien erop toe dat van alle stappen in het proces van databewerking een provenance-spoor wordt bijgehouden.

Organisatie
Het Huygens ING reserveert middelen en tijd voor de instandhouding van de datainfrastructuur en het bijbehorende datamanagement. Het instituut beschikt over een sterke afdeling IT en een snel groeiende afdeling Digitaal Databeheer: beide afdelingen hebben voldoende vaste formatie om de continuïteit van deze activiteiten te garanderen.

De afdeling Digitaal Databeheer heeft op dit moment vijftien medewerkers en beschikt over drie data-officers (Harm Nijboer, Lodewijk Petram en Ronald Sluijter). Deze data-officers hebben ruime onderzoekservaring, analytische vaardigheden, en expertise op het gebied van digitale bronnenkritiek en -analyse. Zij werken nauw samen met onderzoekers om er zeker van te zijn dat de onderzoeksdoelen geenszins worden geschaad door de beoogde interoperabiliteit van data. De afdeling heeft ook een groep van data-curators (vergelijkbaar met onderzoeks-assistenten en informatiespecialisten) die assisteren in taken op het gebied van datamanagement, zoals datavergaring en -invoer, dataconversie en modellering, en de samenstelling en het beheer van vocabulaires, ontologieën en metadata-systemen. Tot slot heeft de afdeling ook een open access/data-adviseur (Milo van de Pol). Zijn voornaamste taak is om onderzoekers te adviseren bij de uitvoering van het open access- en open data-beleid van onderzoeksfinanciers en de KNAW. Verder helpt hij bij kwesties omtrent auteursrecht, licenties, academische embargo’s, risicomanagement, etc. De adviseur treedt ook op als verbinder tussen de onderzoekers en de juridische experts van de KNAW en andere (internationale) onderzoeksorganisaties.