

Provenance document, verwerking van datasets door DANS

Als een dataset bij DANS wordt gearchiveerd, zorgen wij ervoor dat de data duurzaam toegankelijk en leesbaar blijft. Om dit te kunnen garanderen zijn verschillende acties nodig. In dit document leest u meer hierover. Als u andere vragen hebt over dataverwerking of het hergebruiken van data kunt u contact opnemen met DANS via info@dans.knaw.nl.

1. Dataverwerkingsprotocol

Nadat data in EASY gedeponereerd zijn, verwerkt een medewerker van DANS de dataset volgens een standaard dataverwerkingsprotocol. Dit protocol heeft als doel veilig te stellen dat data op de langere termijn vindbaar, toegankelijk en inzichtelijk zijn, dit laatste ook zonder tussenkomst van de oorspronkelijke onderzoeker. Een belangrijk onderdeel van het protocol is, indien van toepassing, de controle op privacygevoelige informatie. Dit geldt met name voor surveydata en interviews.

Op grond van dit protocol worden de volgende controles uitgevoerd sinds de oprichting van DANS in 2005:

- Controle op compleetheit van de dataset, zowel wat de gedeponeerde databestanden betreft als bijbehorende documentatiebestanden
- Controle op de leesbaarheid van de bestanden
- Controle op het bestandsformaat. Ook in de toekomst moeten de data- en documentatiebestanden nog geopend en gebruikt kunnen worden. De controle vindt plaats aan de hand van een lijst met voorkeursbestandsformaten
- Controle van de beschrijving van de dataset op compleetheit en juistheid, en verbetering van de presentatie van de beschrijving
- Controle op privacygevoelige informatie, zowel in de bestanden als in de metadata (zie voor de details de volgende paragraaf)
- Controle op de duidelijkheid van de directorystructuur. Als deze niet voldoende inzichtelijk is, wordt de structuur aangepast
- Controle op compleetheit en correctheid van de bestandenlijst bij archeologische datasets die door de depositor is aangeleverd. Dit is een lijst met een korte omschrijving van elk bestand van de dataset.

Een archivaris zal kleine aanpassingen aan de metadata en de bestandsindeling aanbrengen indien dit de duidelijkheid van de dataset ten goede komt. Voor zeer inhoudelijke, ingrijpende wijzigingen zal contact worden opgenomen met de deponerder.

Bij de verwerking van de dataset is het mogelijk dat een archivaris bestanden uit de dataset naar voorkeursbestandsformaten migreert om de duurzaamheid en de toegankelijkheid van de bestanden op de lange termijn te waarborgen. Indien bestanden worden gemigreerd, zullen de gemigreerde bestanden voor gebruik beschikbaar worden gesteld. De oorspronkelijk gedeponeerde bestanden worden daarnaast ook altijd in onaangetaste vorm bij de dataset bewaard.

Alle acties die door een medewerker van DANS op een dataset worden uitgevoerd, worden geregistreerd in een interne administratie.

2. Privacygevoelige data

Als een bestand exacte namen en exacte geboortedata van respondenten bevat, worden deze variabelen verwijderd. Exacte contactgegevens van de respondenten worden ook verwijderd en van postcodes worden alleen de cijfers gehandhaafd. Exacte beroepsnamen worden ook niet beschikbaar gesteld. Het beroep van respondenten kan wel afgeleid worden uit beroepenclassificaties. In het algemeen geldt dus dat alle identificerende variabelen verwijderd worden. Datasets met privacygevoelige gegevens zijn op die manier alleen in een geanonimiseerd formaat beschikbaar. De niet-geanonimiseerde datasets worden overigens wel gearchiveerd, dit met het oog op analysemogelijkheden in de toekomst. Kwalitatieve data, zoals bijvoorbeeld audiofiles met interviews, kunnen niet geanonimiseerd worden, maar zijn wel beschikbaar omdat de geïnterviewde officieel toestemming heeft gegeven voor gebruik van het interview voor wetenschappelijk onderzoek.

3. Persistent Identifier

Bij archivering wordt aan elke dataset een automatisch gegenereerde Persistent Identifier toegevoegd, bestaande uit een combinatie van cijfers en letters. Deze identifier kan in publicaties over analyses op de dataset opgenomen worden bij de referenties naar de dataset. De Persistent Identifier garandeert de blijvende vindbaarheid van de dataset, ondanks wijzigende internetadressen.

4. Toegang tot de dataset

De toegangsrechten voor gebruikers om de bestanden van een dataset te mogen zien of downloaden zijn door de depositor bepaald bij het aanleveren van de beschrijving van de dataset. DANS neemt deze toegangsrechten over bij het archiveren en toegankelijk maken van de data.

5. Aanvullende informatie

Lees meer over de voorkeursbestandsformaten bij DANS op:
<http://www.dans.knaw.nl/nl/deponeren/toelichting-data-deponeren>