Breakoutsessie Datamanagementplannen

Last of lust voor de onderzoeker?

Fieke Schoots, UB Leiden   Surf Masterclass   4 april 2014

Universiteit Leiden
The Netherlands

Discover the world at Leiden University

Onderzoeker wil met DMP

Onderlinge afspraken goed vastleggen :

- Eigendom
- Verantwoordelijkheden
- Omgang met data

Tijd besparen :

- Formats
- Backup en versiebeheer
- Metadata & documentatie standaardiseren

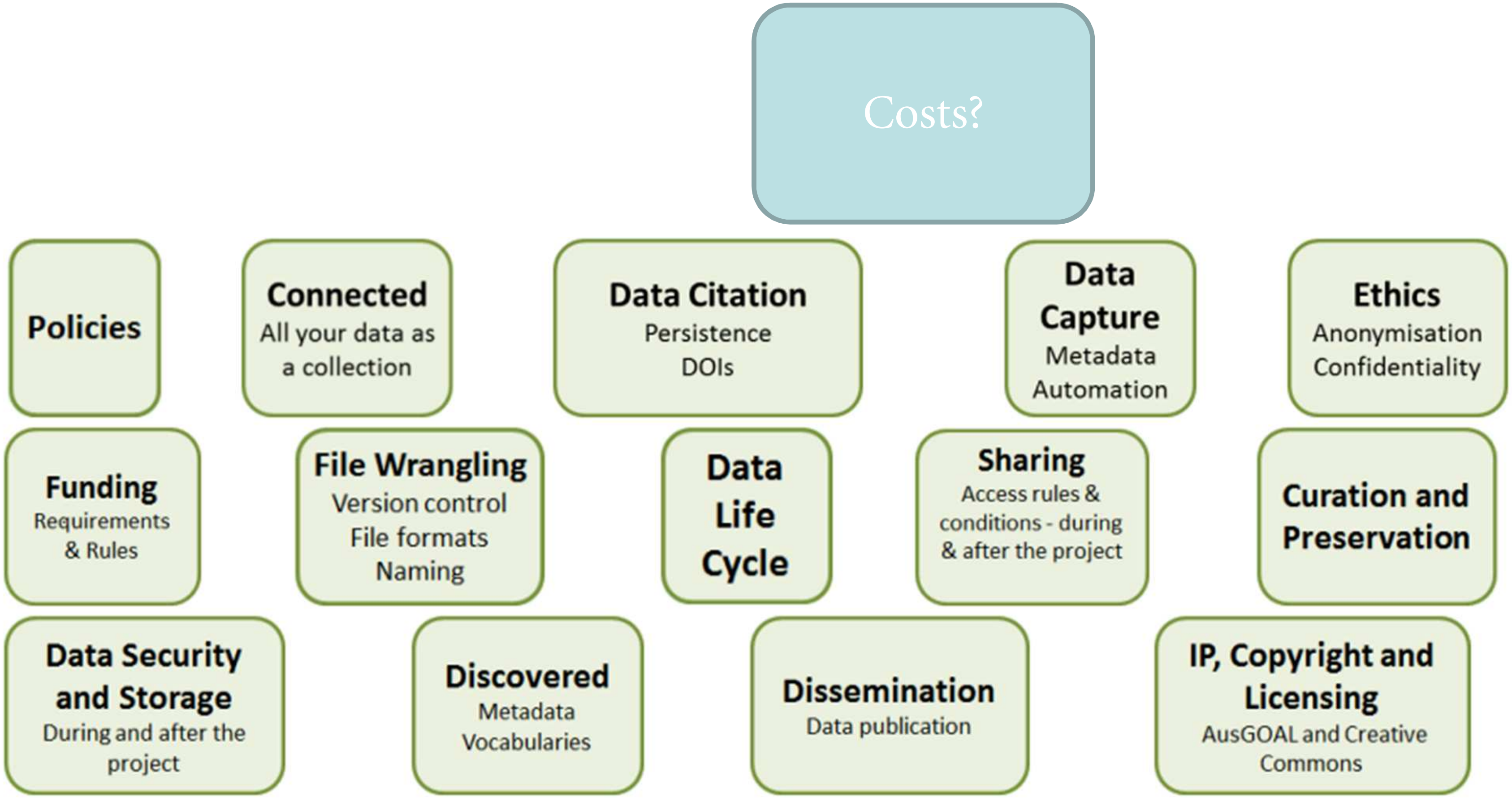
Goede opslagvoorzieningen :

- Qua omvang & veiligheid (ethisch, juridisch)
- Data delen
- Credits

Oz. moet vragen beantwoorden over :

1. Beleid
2. Techniek
3. Onderzoeksmethode en best practice
4. Documentatie (metadata)
5. Juridische & ethische kwesties
6. Toekomst : publicatie, archivering
7. Kosten

Elements of Data Management Plans



Costs?

Policies

Connected
All your data as a collection

Data Citation
Persistence
DOIs

Data Capture
Metadata
Automation

Ethics
Anonymisation
Confidentiality

Funding
Requirements & Rules

File Wrangling
Version control
File formats
Naming

Data Life Cycle

Sharing
Access rules & conditions - during & after the project

Curation and Preservation

Data Security and Storage
During and after the project

Discovered
Metadata
Vocabularies

Dissemination
Data publication

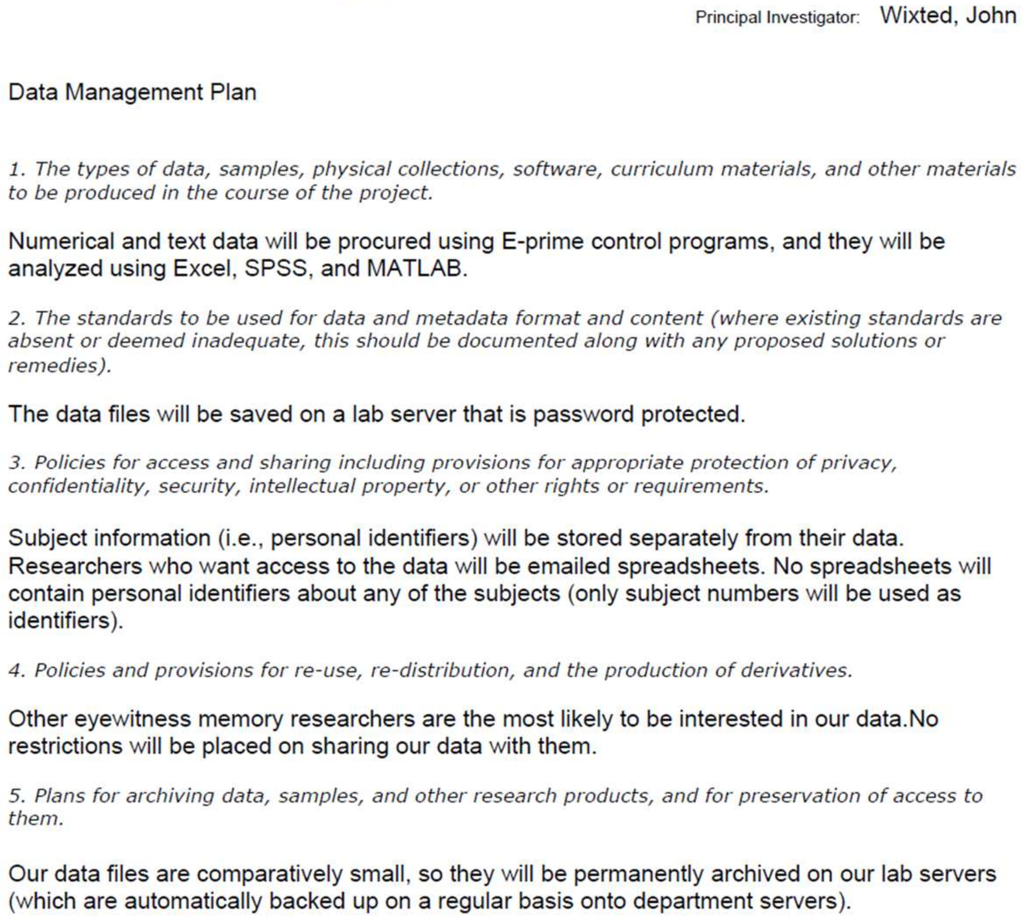
IP, Copyright and Licensing
AusGOAL and Creative Commons

http://ands.org.au/resource/data-management-planning.html

Last ?

- Administratieve klus
- Mijn onderzoek is uniek, er is geen geschikt template
- Ik weet niet wat de regels / gewoonten / standaarden zijn
- Ik weet niet wanneer ik aan voorwaarden voldoe (veiligheid data, open access)?
- Waar kan ik informatie krijgen?
- Voor wie doe ik het? Waar blijft het? (papieren exercitie?)

De kortste weg….



Principal Investigator: Wixted, John

Data Management Plan

1. The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project.

Numerical and text data will be procured using E-prime control programs, and they will be analyzed using Excel, SPSS, and MATLAB.

2. The standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies).

The data files will be saved on a lab server that is password protected.

3. Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements.

Subject information (i.e., personal identifiers) will be stored separately from their data. Researchers who want access to the data will be emailed spreadsheets. No spreadsheets will contain personal identifiers about any of the subjects (only subject numbers will be used as identifiers).

4. Policies and provisions for re-use, re-distribution, and the production of derivatives.

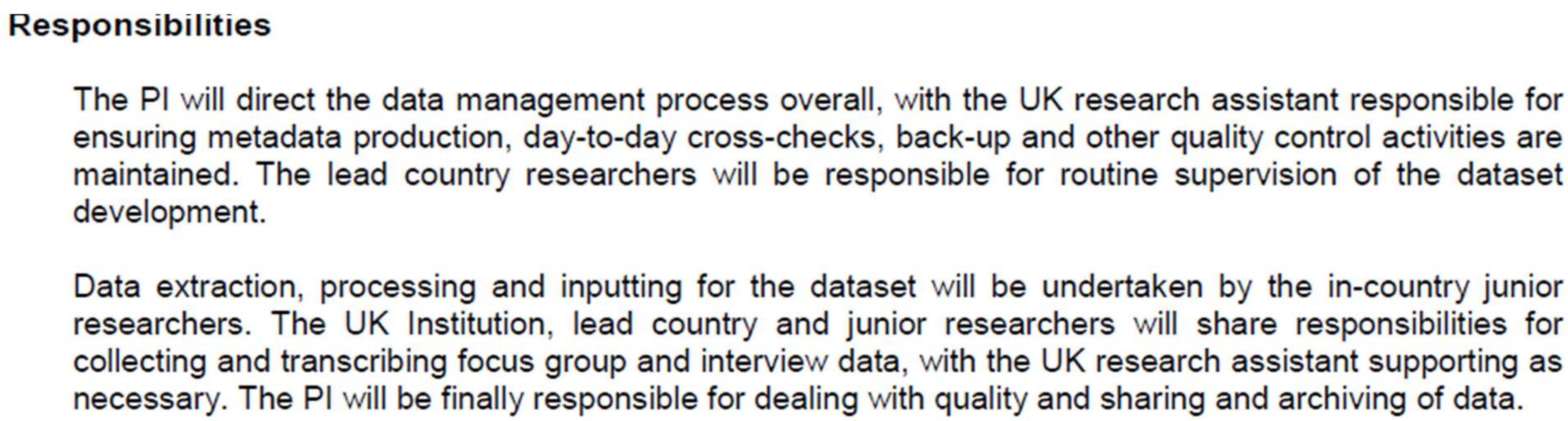
Other eyewitness memory researchers are the most likely to be interested in our data.No restrictions will be placed on sharing our data with them.

5. Plans for archiving data, samples, and other research products, and for preservation of access to them.

Our data files are comparatively small, so they will be permanently archived on our lab servers (which are automatically backed up on a regular basis onto department servers).

http://rci.ucsd.edu/_files/DMP%20Example%20Wixted.pdf

Voorbeeld : verantwoordelijkheden



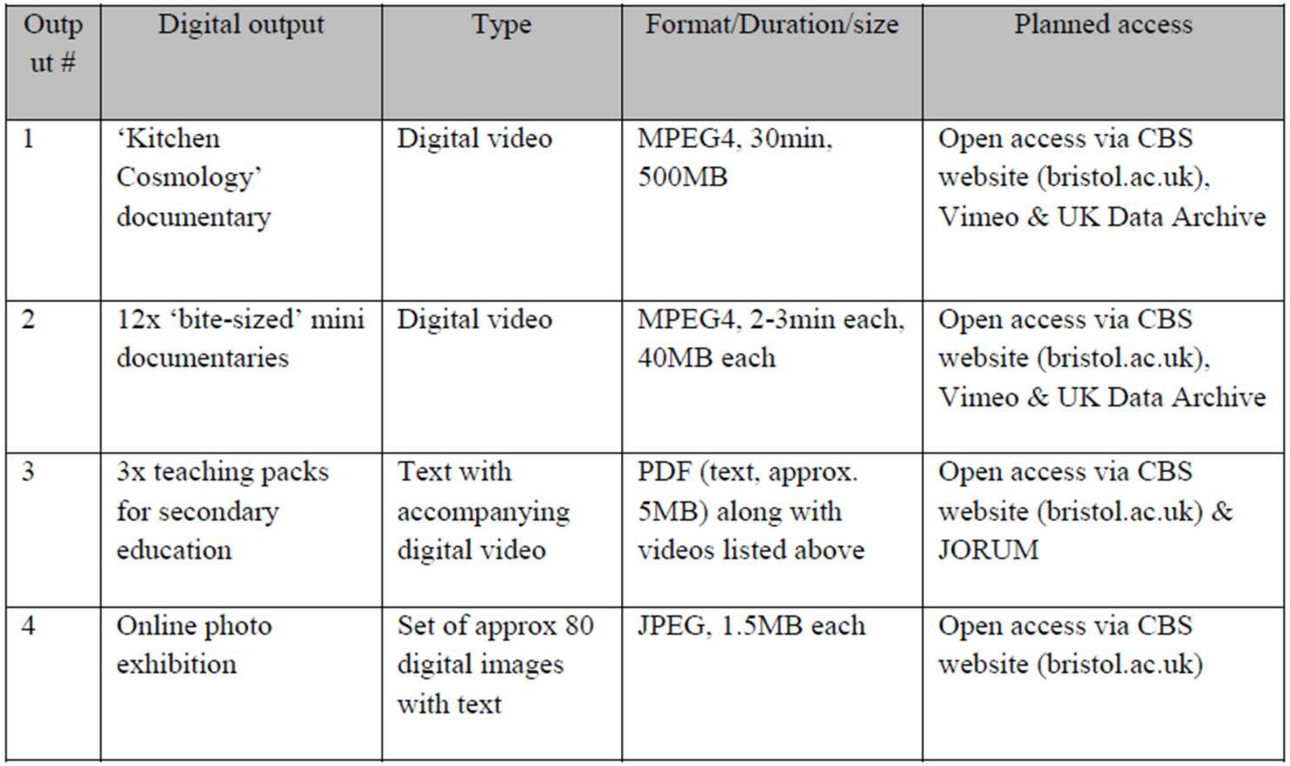
Responsibilities

The PI will direct the data management process overall, with the UK research assistant responsible for ensuring metadata production, day-to-day cross-checks, back-up and other quality control activities are maintained. The lead country researchers will be responsible for routine supervision of the dataset development.

Data extraction, processing and inputting for the dataset will be undertaken by the in-country junior researchers. The UK Institution, lead country and junior researchers will share responsibilities for collecting and transcribing focus group and interview data, with the UK research assistant supporting as necessary. The PI will be finally responsible for dealing with quality and sharing and archiving of data.

From Leeds RoaDMaP Engineering training handbook available at: http://library.leeds.ac.uk/info/377/roadmap/123/roadmap_events/2
http://www.dcc.ac.uk/sites/default/files/documents/adocs/Leeds-RoaDMaP-DMPs.pdf

Voorbeeld : formats (1/3)

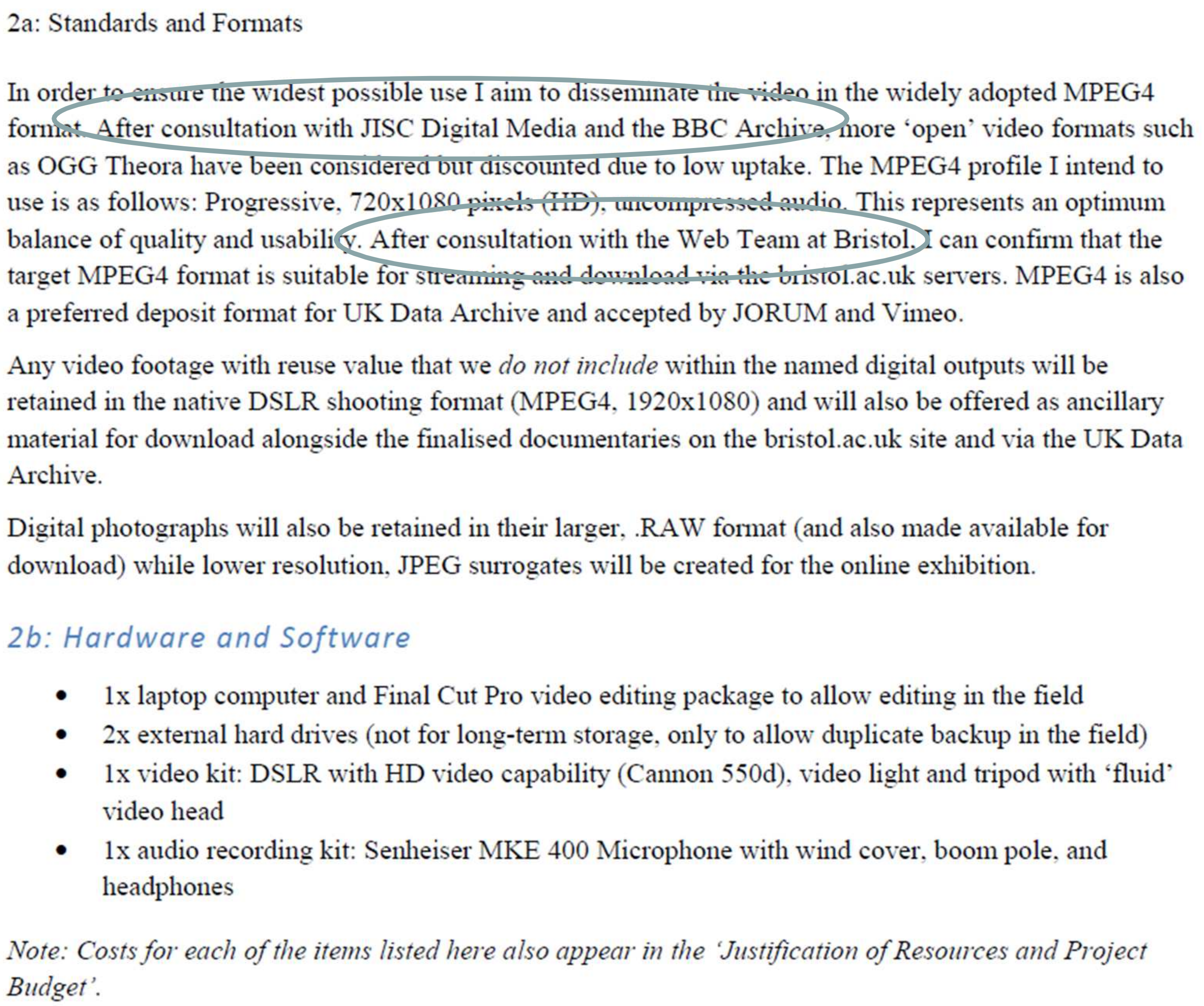


Outp ut #	Digital output	Type	Format/Duration/size	Planned access
1	'Kitchen Cosmology' documentary	Digital video	MPEG4, 30min, 500MB	Open access via CBS website (bristol.ac.uk), Vimeo & UK Data Archive
2	12x 'bite-sized' mini documentaries	Digital video	MPEG4, 2-3min each, 40MB each	Open access via CBS website (bristol.ac.uk), Vimeo & UK Data Archive
3	3x teaching packs for secondary education	Text with accompanying digital video	PDF (text, approx. 5MB) along with videos listed above	Open access via CBS website (bristol.ac.uk) & JORUM
4	Online photo exhibition	Set of approx 80 digital images with text	JPEG, 1.5MB each	Open access via CBS website (bristol.ac.uk)

Arts & Humanities Research Council
Applicant Dr. Rita Langer, Department of Religion and Theology, University of Bristol, 6/3/2013
http://data.bris.ac.uk/research/planning/files/2013/08/data.bris-AHRC-example-Technical-Plan-v2.pdf

Copyright © 2013 University of Bristol

Voorbeeld : formats (2/3)



2a: Standards and Formats

In order to ensure the widest possible use I aim to disseminate the video in the widely adopted MPEG4 format. After consultation with JISC Digital Media and the BBC Archive, more 'open' video formats such as OGG Theora have been considered but discounted due to low uptake. The MPEG4 profile I intend to use is as follows: Progressive, 720x1080 pixels (HD), uncompressed audio. This represents an optimum balance of quality and usability. After consultation with the Web Team at Bristol, I can confirm that the target MPEG4 format is suitable for streaming and download via the bristol.ac.uk servers. MPEG4 is also a preferred deposit format for UK Data Archive and accepted by JORUM and Vimeo.

Any video footage with reuse value that we *do not include* within the named digital outputs will be retained in the native DSLR shooting format (MPEG4, 1920x1080) and will also be offered as ancillary material for download alongside the finalised documentaries on the bristol.ac.uk site and via the UK Data Archive.

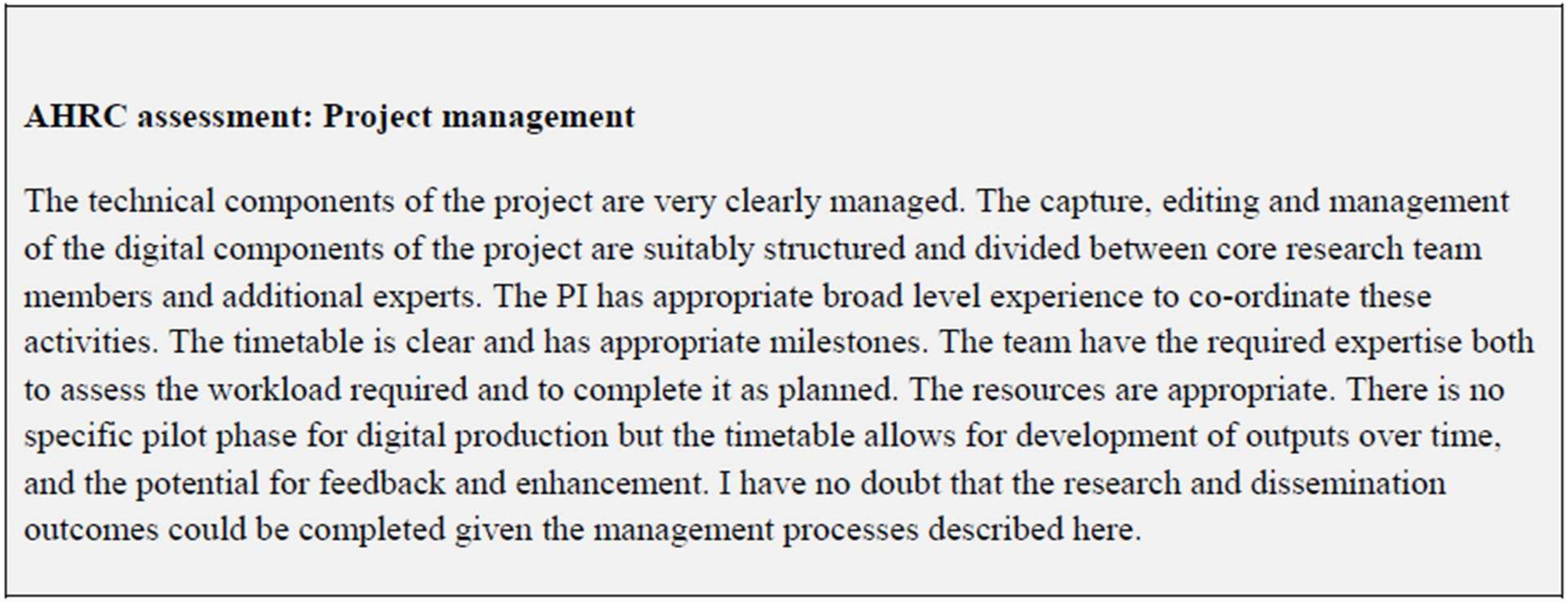
Digital photographs will also be retained in their larger, .RAW format (and also made available for download) while lower resolution, JPEG surrogates will be created for the online exhibition.

2b: Hardware and Software

- 1x laptop computer and Final Cut Pro video editing package to allow editing in the field
- 2x external hard drives (not for long-term storage, only to allow duplicate backup in the field)
- 1x video kit: DSLR with HD video capability (Cannon 550d), video light and tripod with 'fluid' video head
- 1x audio recording kit: Senheiser MKE 400 Microphone with wind cover, boom pole, and headphones

Note: Costs for each of the items listed here also appear in the 'Justification of Resources and Project Budget'.

Voorbeeld formats (3/3)



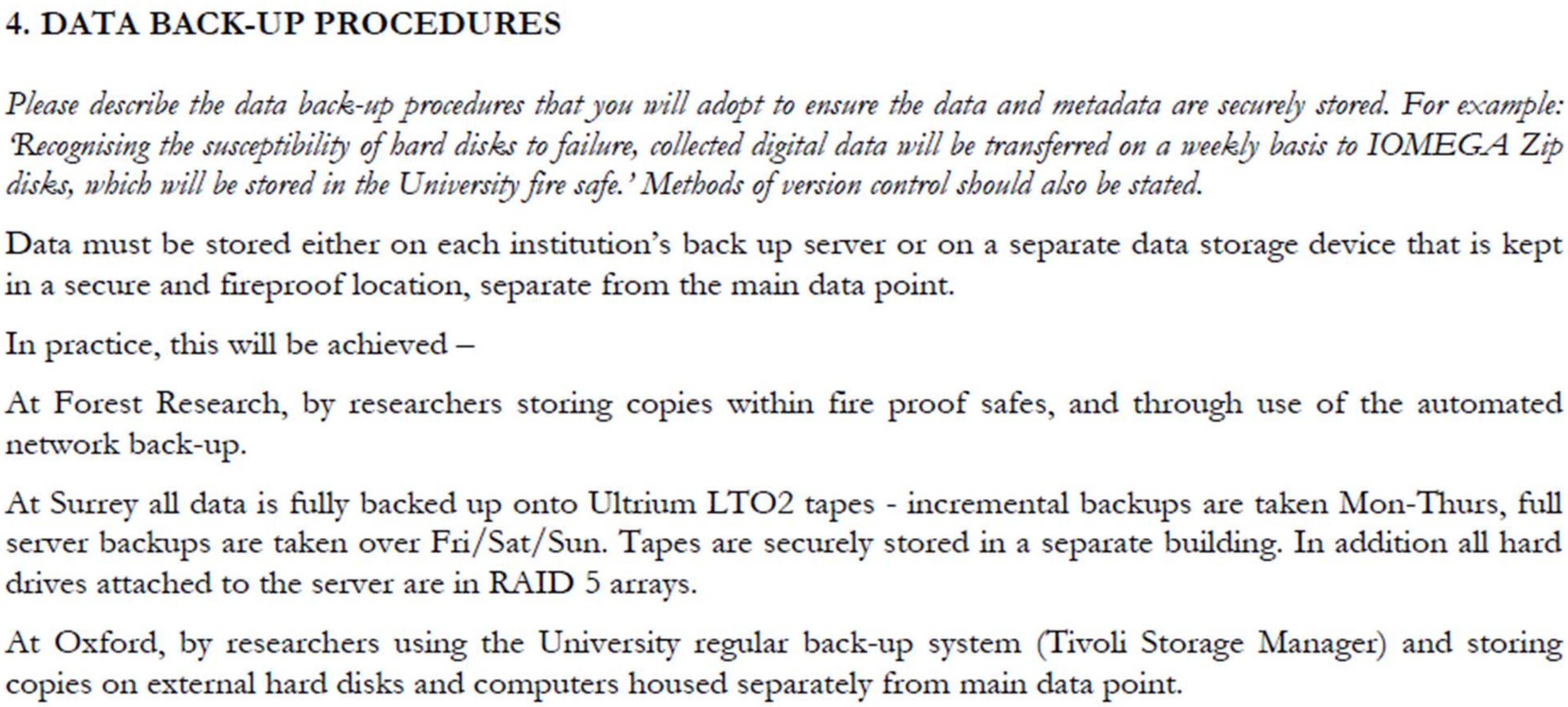
AHRC assessment: Project management

The technical components of the project are very clearly managed. The capture, editing and management of the digital components of the project are suitably structured and divided between core research team members and additional experts. The PI has appropriate broad level experience to co-ordinate these activities. The timetable is clear and has appropriate milestones. The team have the required expertise both to assess the workload required and to complete it as planned. The resources are appropriate. There is no specific pilot phase for digital production but the timetable allows for development of outputs over time, and the potential for feedback and enhancement. I have no doubt that the research and dissemination outcomes could be completed given the management processes described here.

http://www.dcc.ac.uk/sites/default/files/documents/adocs/Leeds-RoaDMaP-DMPs.pdf

Copyright © 2013 University of Bristol

Voorbeeld: backup en versiebeheer



4. DATA BACK-UP PROCEDURES

Please describe the data back-up procedures that you will adopt to ensure the data and metadata are securely stored. For example: 'Recognising the susceptibility of hard disks to failure, collected digital data will be transferred on a weekly basis to IOMEGA Zip disks, which will be stored in the University fire safe.' Methods of version control should also be stated.

Data must be stored either on each institution's back up server or on a separate data storage device that is kept in a secure and fireproof location, separate from the main data point.

In practice, this will be achieved –

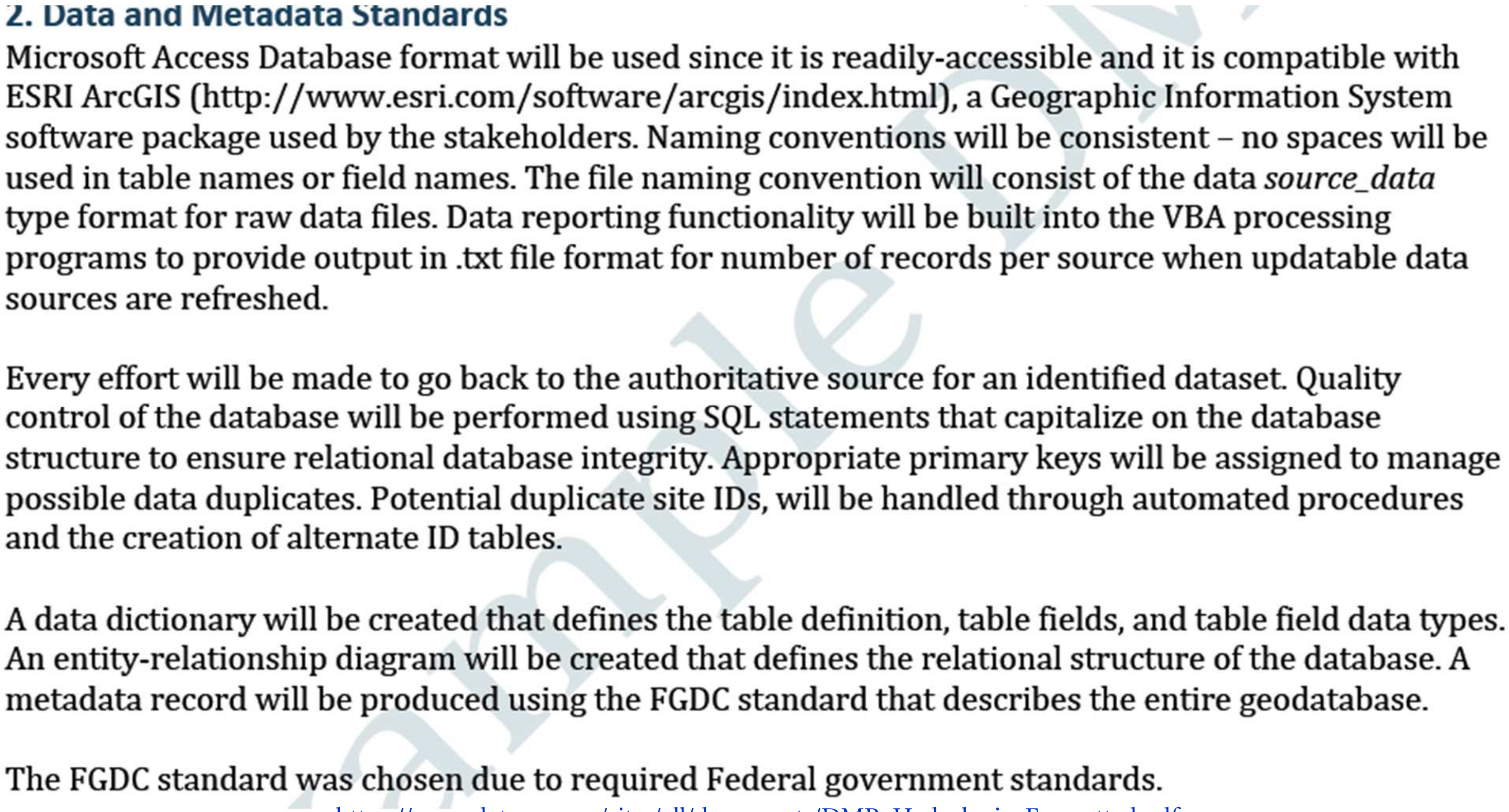
At Forest Research, by researchers storing copies within fire proof safes, and through use of the automated network back-up.

At Surrey all data is fully backed up onto Ultrium LTO2 tapes - incremental backups are taken Mon-Thurs, full server backups are taken over Fri/Sat/Sun. Tapes are securely stored in a separate building. In addition all hard drives attached to the server are in RAID 5 arrays.

At Oxford, by researchers using the University regular back-up system (Tivoli Storage Manager) and storing copies on external hard disks and computers housed separately from main data point.

http://relu.data-archive.ac.uk/media/25706/exampledmp_fr.pdf
© University of Essex [example from http://relu.data-archive.ac.uk/data-sharing/planning/examples]

Voorbeeld : metadata & documentatie



2. Data and Metadata Standards

Microsoft Access Database format will be used since it is readily-accessible and it is compatible with ESRI ArcGIS (http://www.esri.com/software/arcgis/index.html), a Geographic Information System software package used by the stakeholders. Naming conventions will be consistent – no spaces will be used in table names or field names. The file naming convention will consist of the data *source_data* type format for raw data files. Data reporting functionality will be built into the VBA processing programs to provide output in .txt file format for number of records per source when updatable data sources are refreshed.

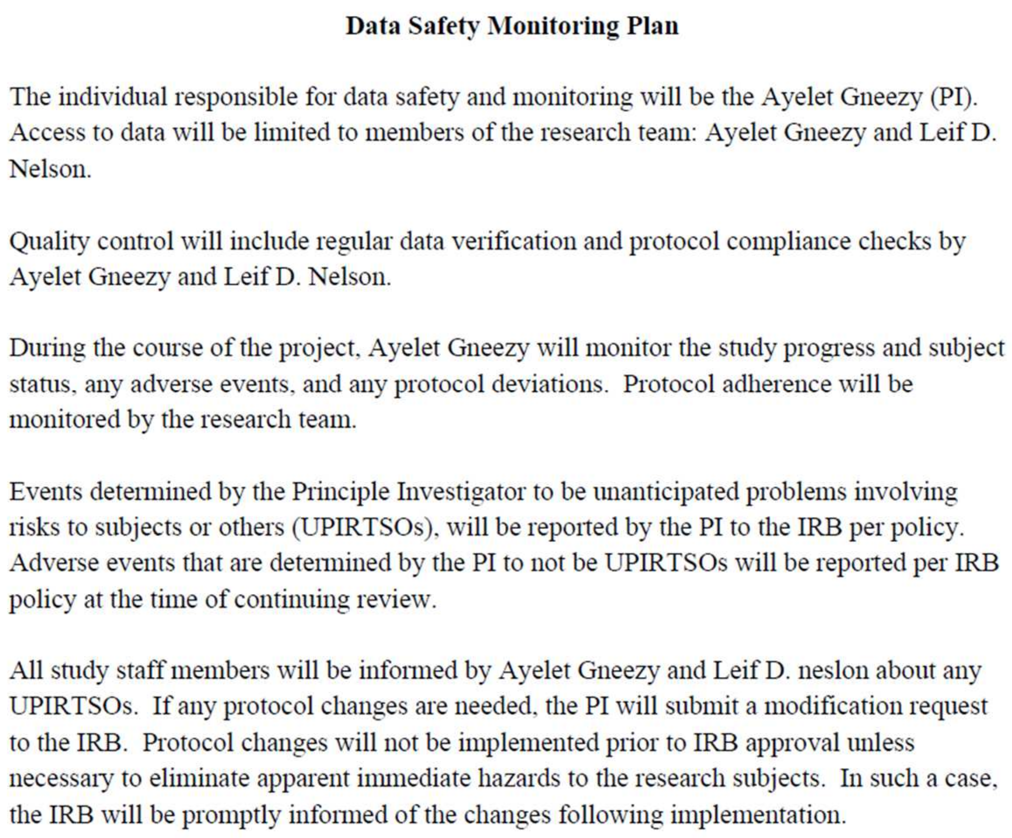
Every effort will be made to go back to the authoritative source for an identified dataset. Quality control of the database will be performed using SQL statements that capitalize on the database structure to ensure relational database integrity. Appropriate primary keys will be assigned to manage possible data duplicates. Potential duplicate site IDs, will be handled through automated procedures and the creation of alternate ID tables.

A data dictionary will be created that defines the table definition, table fields, and table field data types. An entity-relationship diagram will be created that defines the relational structure of the database. A metadata record will be produced using the FGDC standard that describes the entire geodatabase.

The FGDC standard was chosen due to required Federal government standards.

https://www.dataone.org/sites/all/documents/DMP_Hydrologic_Formatted.pdf
Example DMP – Rio Grande hydrology. © DataONE 2011

Voorbeeld : data veiligheid



Data Safety Monitoring Plan

The individual responsible for data safety and monitoring will be the Ayelet Gneezy (PI). Access to data will be limited to members of the research team: Ayelet Gneezy and Leif D. Nelson.

Quality control will include regular data verification and protocol compliance checks by Ayelet Gneezy and Leif D. Nelson.

During the course of the project, Ayelet Gneezy will monitor the study progress and subject status, any adverse events, and any protocol deviations. Protocol adherence will be monitored by the research team.

Events determined by the Principle Investigator to be unanticipated problems involving risks to subjects or others (UPIRTSOs), will be reported by the PI to the IRB per policy. Adverse events that are determined by the PI to not be UPIRTSOs will be reported per IRB policy at the time of continuing review.

All study staff members will be informed by Ayelet Gneezy and Leif D. neslon about any UPIRTSOs. If any protocol changes are needed, the PI will submit a modification request to the IRB. Protocol changes will not be implemented prior to IRB approval unless necessary to eliminate apparent immediate hazards to the research subjects. In such a case, the IRB will be promptly informed of the changes following implementation.

http://rci.ucsd.edu/_files/DMP%20Example%20Ayelet%20Gneezy.pdf
University of San Diego [example from http://rci.ucsd.edu/data-curation/examples.html]

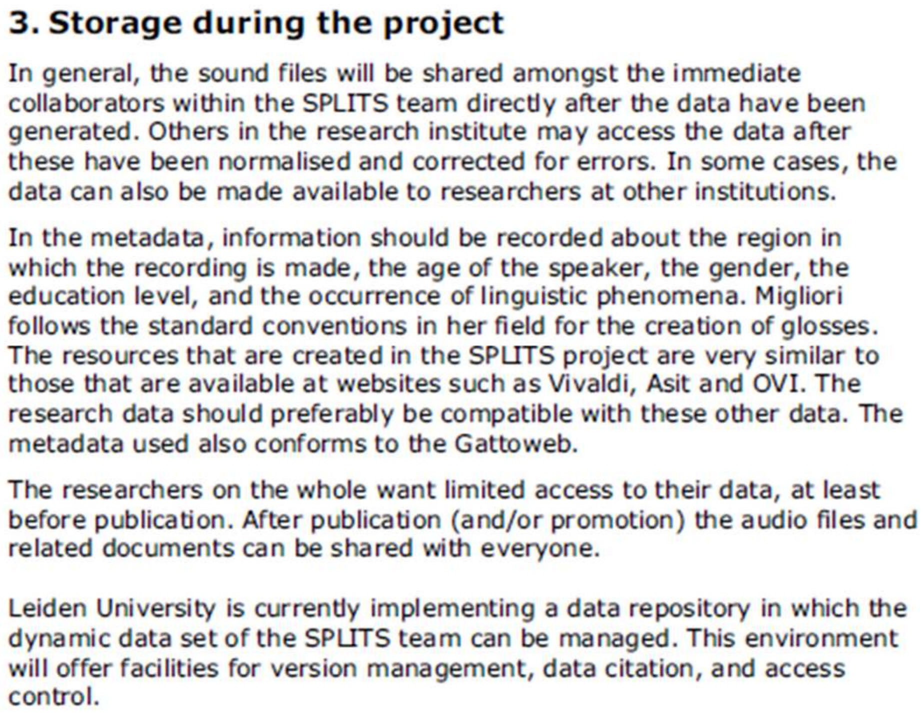
Data veiligheid : EU ethical review

Voorbeelden van aan te leveren documenten in kader van ethical review (FP7) :

"-Applicants must obtain from the data controller of their institution, **written approvals for the technical data protection procedures** that will be implemented in the project. These approvals must demonstrate **compliance of the data protection processes with the European legal framework**. Copies of these approvals must be forwarded to the European Commission prior to the commencement of the related studies. If requested by their institution's data controller, **applicants must consider obtaining approvals/opinions/authorizations from their national data protection authorities** for the intended data collection and processing (http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/others/2006-07-03-vademecum.doc).
If requested, copies of these documents must also be provided to the European Commission.
In cases where the host institution does not have a dedicated data protection officer, and only then, the applicant must contact the national/regional data protection authorities.."

"In order to protect the participants, and to avoid the risk of mosaic effect, the applicant must ensure, for each of the data types collected, that the data are fully anonymized at the stage of collection and that no links are available to go back to the identity of the potential respondents. The applicant must confirm that no respondents' identity will ever be released. **A data processing and merging assessment document must be provided to the ERCEA. A risk mitigation policy needs to be in place."**

Voorbeeld : data delen



3. Storage during the project

In general, the sound files will be shared amongst the immediate collaborators within the SPLITS team directly after the data have been generated. Others in the research institute may access the data after these have been normalised and corrected for errors. In some cases, the data can also be made available to researchers at other institutions.

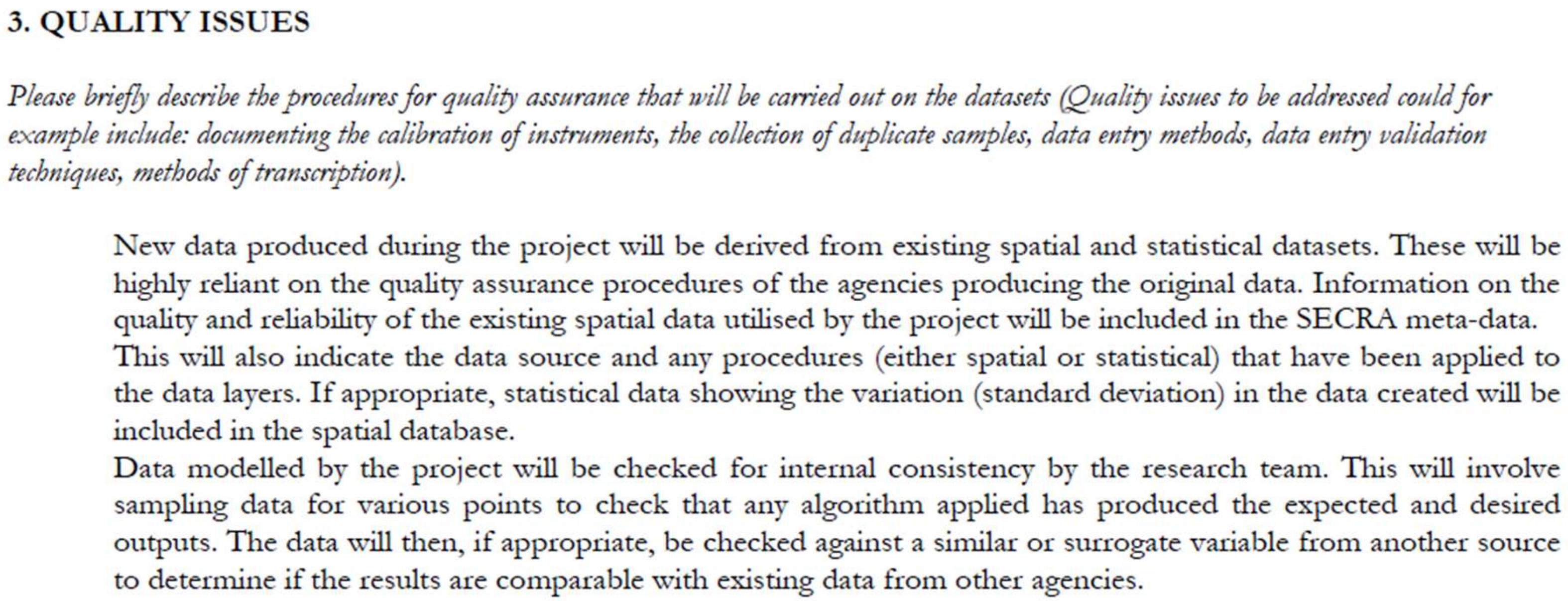
In the metadata, information should be recorded about the region in which the recording is made, the age of the speaker, the gender, the education level, and the occurrence of linguistic phenomena. Migliori follows the standard conventions in her field for the creation of glosses. The resources that are created in the SPLITS project are very similar to those that are available at websites such as Vivaldi, Asit and OVI. The research data should preferably be compatible with these other data. The metadata used also conforms to the Gattoweb.

The researchers on the whole want limited access to their data, at least before publication. After publication (and/or promotion) the audio files and related documents can be shared with everyone.

Leiden University is currently implementing a data repository in which the dynamic data set of the SPLITS team can be managed. This environment will offer facilities for version management, data citation, and access control.

Data Management Plan for the VIDI research group Splitting and clustering grammatical information (SPLITS), Italian Department of Leiden University (4/11/2011) ; zie : http://datasupport.researchdata.nl/start-de-cursus/ii-planfase/datamanagementplanning/

Voorbeeld : kwaliteit



3. QUALITY ISSUES

Please briefly describe the procedures for quality assurance that will be carried out on the datasets (Quality issues to be addressed could for example include: documenting the calibration of instruments, the collection of duplicate samples, data entry methods, data entry validation techniques, methods of transcription).

New data produced during the project will be derived from existing spatial and statistical datasets. These will be highly reliant on the quality assurance procedures of the agencies producing the original data. Information on the quality and reliability of the existing spatial data utilised by the project will be included in the SECRA meta-data. This will also indicate the data source and any procedures (either spatial or statistical) that have been applied to the data layers. If appropriate, statistical data showing the variation (standard deviation) in the data created will be included in the spatial database.

Data modelled by the project will be checked for internal consistency by the research team. This will involve sampling data for various points to check that any algorithm applied has produced the expected and desired outputs. The data will then, if appropriate, be checked against a similar or surrogate variable from another source to determine if the results are comparable with existing data from other agencies.

http://relu.data-archive.ac.uk/media/25709/exampledmp_seira.pdf

© University of Essex [example from http://relu.data-archive.ac.uk/data-sharing/planning/examples]

Data curation profiles

“a tool for librarians, and others, who want or need to gather information about data generated and used in research that may be published, shared, and archived for re-use and dissemination.”

“At an individual level, the Data Curation Profile:

- provides a structure for conducting a data interview between an information professional and a researcher or research group
- provides a means for a researcher or a research group to thoughtfully consider their needs for the data beyond its immediate use”

http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1032&context=dcp

Geen dmp maar interviews : ‘data narrative’

Pilot Leiden : DMP sociale wetenschappen

Pilot o.a. op basis van template van ZonMW en Wageningen.

Conclusie⇨ behoefte aan informatie over

- Wettelijke richtlijnen
  VSNU, CBP, Gedragscode Gezondheidsonderzoek, etc.
- Kennisdelen niveau instituut en universiteit (best practices, vakrichtlijnen, UL beleid Informatie Beveiliging, etc.)
- Dataopslagmogelijkheden en -kosten
- Ondersteuning/expertise binnen de Universiteit en daarbuiten
  UBL (auteursrecht, datamanagement, VRE's)
  Facultair bv.: Graduate Schools
  Landelijk bv.: DANS, 3TU.Datacentrum, etc.

Pilot Leiden : DMP sociale wetenschappen

Aanbevelingen:

- Bestaande sjablonen voldoen niet 100%
- Invullen moet ten goede komen aan de onderzoeker en niet teveel extra moeite kosten
  - Subsidieaanvraag
  - Bijdragen aan de onderzoeks-kwaliteitscyclus
  - Ondersteuningsvraag concreet maken

Pilot Leiden : DMP sociale wetenschappen

Inhoud DMP template

- Thema’s uit bestaande lijsten
- Vraaggestuurd (‘branching’)
- Algemeen deel én instituutspecifiek deel)
- Mix van open vragen en checklist
- Opslag online?
  (opslagbehoeftes ; overzicht data productie)

Lust !

- Informatie (regels, procedures etc.) goed toegankelijk
- Kant-en-klare tekstsuggesties (kan alleen per instelling?)
- Specificaties van opslagvoorzieningen, beleid informatie beveiliging etc.
- Delen ‘research practices’ in instituut : beschrijvingen van methodes van verzamelen, analyses, delen van data
- Voorbeelden van datamanagementplannen beschikbaar stellen (registratie?)

“The DMP is not a fixed document, it evolves and gains more precision and substance during the lifespan of the project. [...] New versions of the DMP should be created whenever important changes to the project occur due to inclusion of new data sets, changes in consortium policies or external factors”. (http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

- En….. ?

Bijlagen bij plan

- Selectiecriteria

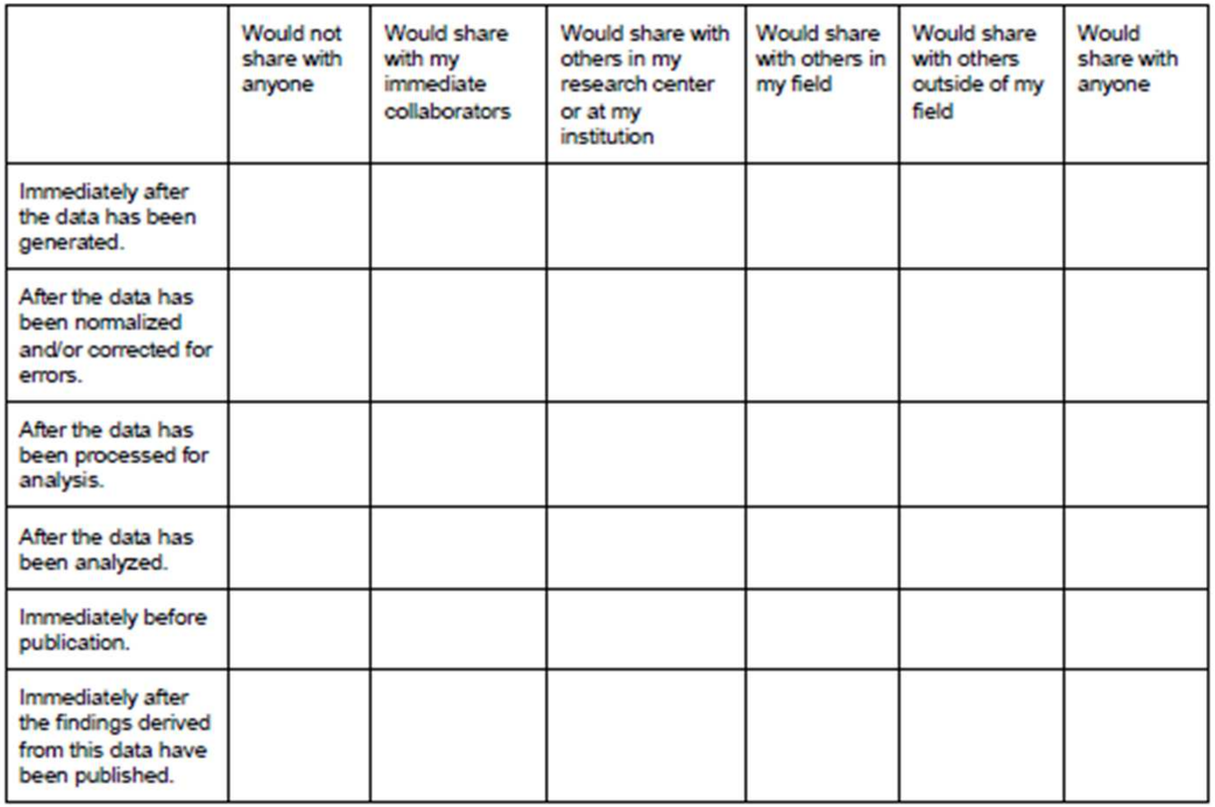
- Schema : met wie wil ik in welk stadium mijn data delen? (uit Data Curation Profiles Toolkit)

- Voorbeelden van folderstructuren en naamgevingsconventies (Datamanagementplan Wageningen)

- Kosten

- Etc.

Data delen schema

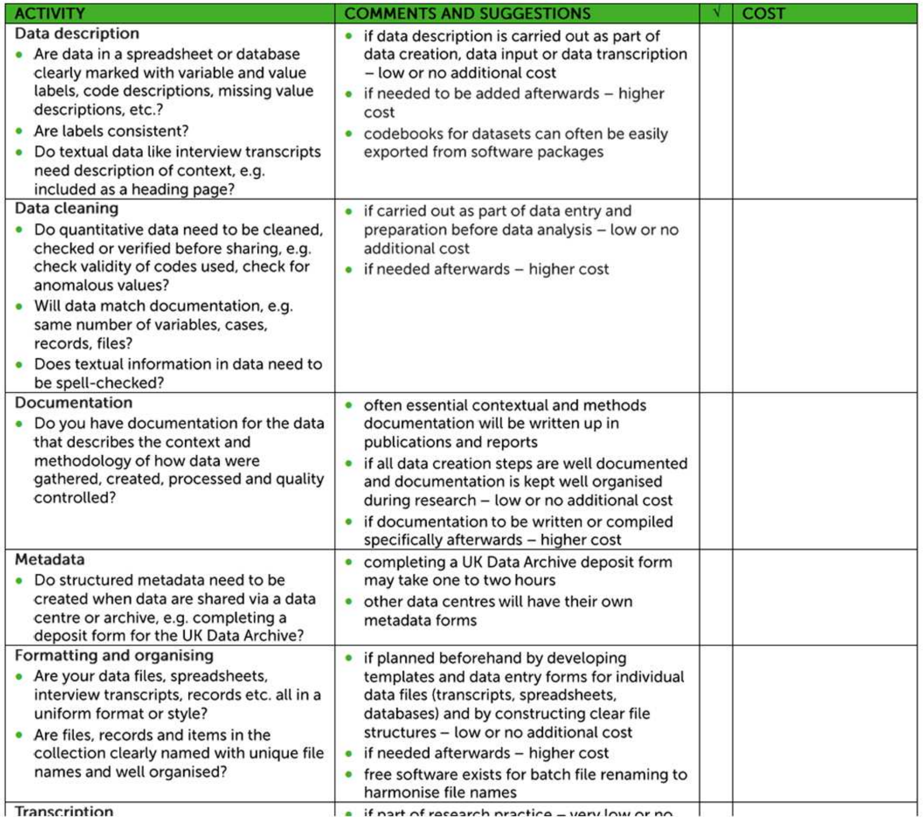


	Would not share with anyone	Would share with my immediate collaborators	Would share with others in my research center or at my institution	Would share with others in my field	Would share with others outside of my field	Would share with anyone
Immediately after the data has been generated.						
After the data has been normalized and/or corrected for errors.						
After the data has been processed for analysis.						
After the data has been analyzed.						
Immediately before publication.						
Immediately after the findings derived from this data have been published.						

Based on : Interview worksheet, Jake Carlson, Purdue University Libraries / Distributed Data Curation Center,

http://ecommons.cornell.edu/bitstream/1813/29064/23/DCP_Interview_Worksheet_v1_DataStaR.pdf

Kosten



ACTIVITY	COMMENTS AND SUGGESTIONS	√	COST
Data description • Are data in a spreadsheet or database clearly marked with variable and value labels, code descriptions, missing value descriptions, etc.? • Are labels consistent? • Do textual data like interview transcripts need description of context, e.g. included as a heading page?	• if data description is carried out as part of data creation, data input or data transcription – low or no additional cost • if needed to be added afterwards – higher cost • codebooks for datasets can often be easily exported from software packages		
Data cleaning • Do quantitative data need to be cleaned, checked or verified before sharing, e.g. check validity of codes used, check for anomalous values? • Will data match documentation, e.g. same number of variables, cases, records, files? • Does textual information in data need to be spell-checked?	• if carried out as part of data entry and preparation before data analysis – low or no additional cost • if needed afterwards – higher cost		
Documentation • Do you have documentation for the data that describes the context and methodology of how data were gathered, created, processed and quality controlled?	• often essential contextual and methods documentation will be written up in publications and reports • if all data creation steps are well documented and documentation is kept well organised during research – low or no additional cost • if documentation to be written or compiled specifically afterwards – higher cost		
Metadata • Do structured metadata need to be created when data are shared via a data centre or archive, e.g. completing a deposit form for the UK Data Archive?	• completing a UK Data Archive deposit form may take one to two hours • other data centres will have their own metadata forms		
Formatting and organising • Are your data files, spreadsheets, interview transcripts, records etc. all in a uniform format or style? • Are files, records and items in the collection clearly named with unique file names and well organised?	• if planned beforehand by developing templates and data entry forms for individual data files (transcripts, spreadsheets, databases) and by constructing clear file structures – low or no additional cost • if needed afterwards – higher cost • free software exists for batch file renaming to harmonise file names		
Transcription	• if part of research practice – very low or no		

http://www.data-archive.ac.uk/media/247429/costingtool.pdf

4C: Collaboration to Clarify the Costs of Curation:
http://www.4cproject.eu/ en http://www.4cproject.eu/community-resources/outputs-and-deliverables/d3-1-evaluation-of-cost-models-and-needs-gaps-analysis ?

Stelling 1

Een DMP moet zo kort mogelijk zijn, om de onderzoeker niet onnodig te belasten.

Stelling 2

Veel instellingen maken een eigen template, maar zouden beter kunnen bevorderen dat vakgenoten (over instellingen heen) disciplinegebonden templates maken.

Stelling 3

Er moet een online tool ontwikkeld worden voor de Nederlandse situatie (naar het voorbeeld van DMP Online of DMP tool) waarin de onderzoekers een datamanagementplan kunnen opstellen. Wie zou dat moeten doen?

Stelling 4

Het is de taak van de instelling om te investeren in de kennis en de infrastructuur die nodig is om te voldoen aan de nieuwe eisen van de subsidieverstrekkers m.b.t. datamanagement (ZonMW, Horizon 2020). Dit is niet de verantwoordelijkheid van de financier. De instelling moet onderzoekers bijvoorbeeld zelf kunnen helpen bij het maken van adequate kostenberamingen.