Implementation of the Data Seal of Approval

The Data Seal of Approval board hereby confirms that the Trusted Digital repository The Language Archive - Max Planck Institute for Psycholinguistics complies with the guidelines version 1 of 2010 set by the Data Seal of Approval Board.
The afore-mentioned repository has therefore acquired the Data Seal of Approval of 2010 on March 1, 2011.

The Trusted Digital repository is allowed to place an image of the Data Seal of Approval logo corresponding to the guidelines version date on their website. This image must link to this file which is hosted on the Data Seal of Approval website.

Yours sincerely,

The Data Seal of Approval Board

Assessment Information

Guidelines Version:	1 | June 1, 2010
Guidelines Information Booklet:	DSA-booklet_1_June2010.pdf
All Guidelines Documentation:	Documentation
Repository:	The Language Archive - Max Planck Institute for Psycholinguistics
Seal Acquiry Date:	Mar. 01, 2011
For the latest version of the awarded DSA for this repository please visit our website:	http://assessment.datasealofapproval.org/seals/
Previously Acquired Seals:	None
This repository is owned by:	**Max Planck Institute for Psycholinguistics** • Wundtlaan 1 6525XD Nijmegen The Netherlands T +31-24-3521911 F +31-24-3521213 E Paul.Trilsbeek@mpi.nl W http://tla.mpi.nl/

Assessment

1. The data producer deposits the research data in a data repository with sufficient information for others to assess the scientific and scholarly quality of the research data and compliance with disciplinary and ethical norms.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

This guideline cannot be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

The repository is divided into different sub-repositories. There is a part for research that has been conducted by researchers affiliated to the Max Planck Institute for Psycholinguistics, there is a part for DOBES endangered language documentation projects and there are parts for other related projects. There is also a part for non-related researchers or projects. On the basis of the originating institution or organization, a data consumer can make some judgments about the level of trust or about the reputation of the depositor. Depositors within the Max Planck Institute are bound to ethical rules regarding human subject data from the Max Planck Society. Depositors in a DOBES project are bound to ethical rules of the DOBES code of conduct. The archive does not (and cannot) systematically verify whether the data it receives is collected according to these rules.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

2. The data producer provides the research data in formats recommended by the data repository.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

This guideline cannot be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

The repository has a list of accepted file formats. Only these formats are accepted by the ingest tool, which checks for validity of the ingested resources. Other file formats need to be converted, the repository offers advice on how to do this or in some cases does the conversions for the depositor.

http://www.lat-mpi.eu/tools/lamus/manual/apa.html

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

3. The data producer provides the research data together with the metadata requested by the data repository.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

This guideline cannot be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

The data producer is required to provide metadata in the IMDI format. Metadata descriptions are generally created for bundles of resources that belong together, e.g. an audio recording with its transcription. The repository offers tools for the creation of these metadata descriptions and offers training on metadata creation. There is a recommendation for a minimum set of metadata fields that need to be filled in, but this is not enforced. The technical compliance of the submitted metadata to the IMDI schema is validated during ingest.

http://www.mpi.nl/imdi
http://www.lat-mpi.eu/tools/imdi

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

4. The data repository has an explicit mission in the area of digital archiving and promulgates it.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

This guideline can be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

We have an explicit mission to archive language resources from all around the world, both collected by associated researchers as well as researchers who are not affiliated with us. We promote this mission as much as possible in international conferences an during training courses that we organize ourselves or training courses that we are asked to take part in. The mission goes together with the official possibility to store full copies at two computer centers at different locations for which the president of the Max Planck Society gives an institutional backing of 50 years of bit-stream preservation. We are working on duplicating the archive access framework in those backup locations as well, such that access to the data can be provided even if our institute would cease to exist.

http://www.mpi.nl/research/research-projects/the-language-archive

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

5. The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

This guideline cannot be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

The repository is not a legal entity on its own but is part of the Max Planck Institute for Psycholinguistics which in its turn is not a legal entity of its own but part of the Max-Planck-Gesellschaft zur Förderung der Wissenschaften. e.V. Eingetragener Verein ("registered association") is its legal status. The repository is funded by the MPI for Psycholinguistics, the Max Planck Society (MPG), the Berlin-Brandenburg Academy of Sciences (BBAW) and the Royal Netherlands Academy of Arts and Sciences (KNAW). The repository has agreements with its external depositors about the right to archive the data. The depositors themselves are responsible for compliance with any legal regulations in the area where the data is collected. Where required by national regulations the archive also signs contracts with national/regional institutions. All ethical issues are dealt with by using Codes of Conduct, such as the DOBES Code of Conduct for the DOBES part of the archive. The repository enables the depositors to restrict access to their resources at various levels. All distributed copies elsewhere are stored under the agreement that they are made available under the same access restrictions, if they are made available.

http://www.mpi.nl/DOBES/ethical_legal_aspects
http://www.mpi.nl/DOBES/archive_access/access_procedures

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

6. The data repository applies documented processes and procedures for managing data storage.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

This guideline can be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

Two copies of every resource are stored within the MPI and at least 4 additional copies are stored in different physical locations in Germany. The storage hardware is being replaced at regular intervals to the latest state of the art. Regular checks are performed on archival content to check for file and format integrity. The Sun SAM-FS HSM system that is being used for storage also checks for file integrity upon file access. The repository will have 2 identical archive access setups at the backup sites in Göttingen and Munich, so that in case of an emergency the data can be accessed via one of these sites.

http://www.mpi.nl/DOBES/archive_info/long_term_persistence

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

7. The data repository has a plan for long-term preservation of its digital assets.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

This guideline can be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

Besides the steps mentioned at the previous guideline to take care of the bit stream preservation of the resources, some measures are taken to enhance the chance of future interpretability of the data. The number of accepted file formats is limited, to make future conversions to other formats more feasible. As much as possible open (non-proprietary) file formats are used. For textual resources, XML formats are used whenever possible, to make future interpretation of the files possible even if the tool that was used to create them no longer exists. Text is encoded in Unicode to ensure future interpretability.

http://www.lat-mpi.eu/tools/lamus/manual/apa.html

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

8. Archiving takes place according to explicit work flows across the data life cycle.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

This guideline can be outsourced.

Applicant Entry

Statement of Compliance:

3. In progress: We are in the implementation phase.

Evidence:

There is an abstract standard workflow, but with technological advances there is a large variety of how this is applied. The online archive management tool LAMUS defines a workflow to a certain extent, because no resources can be archived without metadata being present and without a corpus hierarchy being present. The depositor mainly decides what material is being archived; the archive only has technical criteria about file formats and encodings. The depositor determines who can access the material and is also responsible for protecting the privacy of any subjects appearing in the recordings or texts.

There are no formal criteria in place to decide on when to apply data transformations to the current archival formats. More documentation is required to describe various workflow scenarios.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

9. The data repository assumes responsibility from the data producers for access and availability of the digital objects.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

This guideline cannot be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

The archive has signed agreements with external depositors. For DOBES depositors there is the following agreement:
http://www.mpi.nl/DOBES/ethical_legal_aspects/DOBES-daa-v1.pdf
Agreements with other external depositors are based on this.
Depositors within the MPI for Psycholinguistics are contractually obliged to archive their data, so no agreements are necessary with them. All archived resources are available online, the access permissions are defined by the depositors. The repository will have 2 identical archive access setups at the backup sites in Göttingen and Munich, so that in case of an emergency the data can still be accessed via one of these sites.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

10. The data repository enables the users to utilize the research data and refer to them.

Minimum Required Statement of Compliance:

2. Theoretical: We have a theoretical concept.

This guideline cannot be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

The repository provides various ways of utilizing the archived data via online tools as well as by downloading the data in formats commonly used by the research communities. An advanced metadata search utility is provided, as well as a deep search tool for textual content. All metadata can be harvested via the OAI-PMH protocol. Unique persistent identifiers according to the Handle system are provided for each archived object.
http://corpus1.mpi.nl
http://corpus1.mpi.nl/ds/oai2/

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

11. The data repository ensures the integrity of the digital objects and the metadata.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

This guideline cannot be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

MD5 checksums are calculated for all objects and checked periodically. The availability of files on the file system is checked automatically daily. The availability of the archive access tools is checked automatically multiple times a day. The availability of file, web and application servers is monitored continuously. New versions of archived resources can be deposited, in which case the old versions will be moved to a version archive. In the future these old versions will also be made available to the end users but this is currently not yet the case.

More documentation should be written for this guideline.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

The statement of compliance is good, but I am not sure what the last sentence means:

"More documentation should be written for this guideline."

12. The data repository ensures the authenticity of the digital objects and the metadata.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

This guideline cannot be outsourced.

Applicant Entry

Statement of Compliance:

3. In progress: We are in the implementation phase.

Evidence:

The repository in principle makes the original deposited objects available in an unmodified way, if the objects were in one of the accepted file types and encodings. Additionally, lower quality distribution copies of audio and video recordings are made available. New versions of archived resources can be deposited, in which case the old versions will be moved to a version archive. Different versions of the same resource are not compared; we assume the depositor has good reasons for depositing a newer version. A new version of a resource will get a new persistent identifier; the old version will keep the original persistent identifier. Metadata can change if the depositor or archivist sees the need for that, in the case of errors or missing information. Changes to the metadata are currently not logged. All archived objects are linked to their metadata descriptions and are organized in hierarchical (or multi-rooted) tree structures to indicate relationships between objects and sets of objects. The tree structures can change if the depositors decide that this is necessary. The identities of the depositors are checked by means of a login and password when they deposit material online. Provenance metadata as to who made changes to the repository is currently only stored in log files and not shown to the data consumer.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

13. The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

This guideline can be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

The repository supports the OAIS reference model’s tasks and functions, in so far that they are not in conflict with the Live Archives idea:
http://www.mpi.nl/dam-lr/lra-flyer/

We do not create data packages for Ingest, Archiving and Dissemination for example but we treat each archival object separately while maintaining relational links to metadata and other objects.

The data consumer has direct access to the archived objects via the web, provided that access requirements have been met.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

I am accepting this evidence because the guideline refers to "internationally accepted archival standards like OAIS," and not just OAIS, and because the archive is on the whole so responsive to data preservation and access. In general, though, I have the view that there is a very high bar to full compliance with OAIS and that it is a goal that few archives have reached.

14. The data consumer complies with access regulations set by the data repository.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

This guideline cannot be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

Most of the data in the repository is protected; an account is necessary to get access to the data. For some data sets, explicit permission from the depositor is needed. For a large part of the data, the data consumer needs to agree with a code of conduct, which also contains licensing terms. Some corpora have Creative Commons licenses applied to them. If the data consumer does not comply with the access regulations, the only thing that can be practically done is to deny him/her further access and to make the research community aware of the misuse.
http://www.mpi.nl/DOBES/ethical_legal_aspects/

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

15. The data consumer conforms to and agrees with any codes of conduct that are generally accepted in higher education and scientific research for the exchange and proper use of knowledge and information.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

This guideline cannot be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

There are a number of specific codes of conduct that are applicable to parts of the repository, e.g. the DOBES code of conduct. The codes of conduct are in line with generally accepted codes of conduct for research data in the Netherlands. Users need to agree with the codes of conduct before they get access to the data. http://www.mpi.nl/DOBES/ethical_legal_aspects/

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

16. The data consumer respects the applicable licenses of the data repository regarding the use of the research data.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

This guideline cannot be outsourced.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Evidence:

If applicable, the data consumer is made aware of usage restrictions for the data she/he has gotten access to. Generally the usage restrictions are already described in the codes of conduct. For some data, explicit statements need to be made by the data consumer about the usage of the data before he/she gets access. The depositor then decides on whether access is granted or not. In case of misuse, the only thing that can be practically done is to deny the user further access to the repository and to make the research community aware of the misuse.
http://www.mpi.nl/DOBES/ethical_legal_aspects/

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments: