Archive Organization at the MPI

This guide describes how the language resource archive that is maintained at the MPI is organized. The MPI archive includes contributions from various donators and donator groups.

General Framework:

- In a time where we see an enormous increase in language resources, central archives become increasingly important to take care of long-term stability and accessibility.
- Due to these trends it is important that archives have clear organizational principles evident to everybody involved.

1. IMDI Catalogue

Each archive must have a catalogue system with clearly defined categories and the values that these categories can have. For the MPI archive, the usage of the IMDI metadata system that was designed with the help of field linguists and language engineers and developed within European projects is a MUST. To ensure a long lifetime and to make all components available to everyone interested, the manuals and documentation, the schema, the controlled vocabularies and the sources of the tools are all openly available. They will be maintained by the MPI and changes to the vocabularies will only be accepted when they are communicated with the user community. The IMDI infrastructure is built on five years of consensus building and experience and can now be seen as a mature and stable framework.

IMDI spans a domain of linked metadata descriptions and allows for the grouping of resources into hierarchies. IMDI supports distributed scenarios and can be used for cataloguing, browsing, searching and managing language resources of different types. IMDI files are XML files. Fast indexes are created only for searching, etc.

IMDI comes with a whole range of tools: a professional editor, a browser operating on the XML files, transformation to HTML pages, structured and unstructured search, support for Google, a gateway to DublinCore and OLAC metadata services, a TreeCopier, etc.

2. Separation of Physical and Virtual Layer

IMDI allows us to separate the physical storage layer (servers, discs, directories) from the virtual layer. While the physical layer is only relevant for system managers to carry out their tasks, the users are interested in the virtual organization that is determined by linguistic categories and descriptions. The archive managers can take care of linking these two layers and checking consistency. This separation allows all groups to carry out manipulations without affecting each other.

3. Copying and Migrating

All material ingested into the archive will be subject to the same copying and migration processes. The MPI currently has a multi-layered storage infrastructure that immediately creates two copies of every resource that are located at different places within the building. Furthermore, all resources are copied to the two big data centers of the Max-Planck-Society (GWDG Göttingen and RZG Munich) which also apply a double storage strategy. The DOBES archive is also copied to the MPI for Evolutionary Anthropology. These copying actions are based on dynamic protocols, i.e. each resource is available in six (DOBES seven) copies.

The MPI, as well as the data centers, apply a migration strategy according to which all storage components will be exchanged at regular intervals. Both strategies together increase the chances of survival of the data. The Max-Planck-Society is giving a 50 year institutional backing of the preservation of the data on its data center computers.

Peter Wittenburg, MPI for Psycholinguistics, 21. April 2005

Archive Organization at the MPI

4. Ingesting and Consistency

For ingesting new material or modifying existing material, manual operations by the archive managers are currently necessary. Only archive managers are allowed to carry out operations in the archive to guarantee consistency. From May 2005 we will swap over to the LAMUS system (Language Archive Management and Upload System). LAMUS is the gatekeeper which allows authorized users to add new data or modify existing data in a controlled way. Automatic checks will be carried out to guarantee a high degree of consistency and to ensure that old versions will be maintained.
When LAMUS is tested and well-proven, the MPI will open its archive to other donators. On request valuable language resources can be ingested by other donators.

5. Access Management

Access Management (AM) is essential and implements the requirements imposed by ethical and legal considerations. An AM system was developed that is based on the IMDI metadata infrastructure to allow for efficient operations. The responsible donators have the right to define policies (declaration to be signed, usage statements to be specified) and to set rights. They can delegate these rights to other persons for (parts of) their domain of authority. To set rights, the authorized person can select an arbitrary archive node and use simple commands affecting all resources of a certain type under the selected node. To make the administration of access rights feasible, all interaction about getting access will be done electronically.

6. Access Layers

The archivist recognizes the need to not only ensure the long-term survival of data, but also to give access to it and to be able to enrich it. The MPI archives will support the following facilities:

- Metadata Layer: metadata is open and can be used to find resources, XML as well as HTML browsing is supported; also different searching modes are supported;
- Single Resource Access Layer: once resources have been found they can be downloaded or be visualized directly via browser and plug-in technology;
- Sub-Archive Creation Layer: with the help of the TreeCopying tool, a whole sub-tree of the archive, including metadata and resources, can be downloaded; this can be used to set up a new archive at another location;
- Web-Exploitation Layer: a framework of web-based exploitation tools for working with complex multimedia annotations, lexica and other linguistic texts which allows you to select several different resources and work on them;

7. Advisory Boards

For the DOBES archive an advisory board of well respected field linguists has been formed. This will become active in case of linguistically related problems of all sorts such as access conflicts. For the whole archive at the MPI an Archive Advisory Board will be established which will also include archivists and technologists. This board will be responsible for the continuity of the archive in all respects.