Managing very large Multimedia Archives and their Integration into Federations

Daan Broeder, Eric Auer, Marc Kemps-Snijders, Han Sloetjes,
Peter Wittenburg, Claus Zinn[1]

[1] Max-Planck-Institute for Psycholinguistics, Nijmegen, Netherlands

1 Introduction

While the natural sciences are used to deal with terabytes and even petabytes of data, such dimensions are new in the domain of linguistics resp. in the humanities. The reasons for this are manifold; one certainly was the lack of high resolution and high frequency measurement equipment to record real-life situations. Only in the last decade the situation has changed dramatically. There is a long tradition in experimentally oriented institutes such as at the MPI for Psycholinguistics in working with time series data gathered in experimental laboratories (speech, eye tracking, gesture, etc) and in manipulating such signals with the help of computers. However, the amount of data was in the order of mega- or gigabytes, since the data recording was limited in terms of recording length and resolution both in space and time. However, since the beginning of the 90-ies a thorough digitization strategy was applied also for observational data as it is gathered in endangered language documentation, mother-child interaction, gesture and sign-language studies, for example. In these observation studies recording times exceeding one hour are not exceptional. Of course audio is sampled at hifi frequency in good quality and in stereo resulting in 700 kb per hour and in particular video recordings are demanding. Dependent on the used video codec Gigabytes per hour are generated. Since in parallel also in the humanities new devices such as brain imaging recorders have been introduced, large amounts of data are produced also in experimental studies. This resulted in a digital archive containing now about 35 Terabyte with an annual increase of about 8 Terabyte. The effort needed to manage very large databases is not only correlated with the size in bytes, but it is in particular correlated with its internal complexity. While large holdings in the domain of natural sciences in general have a comparatively simple canonical structure, there is a high level of complexity in humanities holdings. There are a lot of different file types with highly different internal structure and complex semantics, there are various types of relations between the resources and the sheer number of individual resources is incredibly high. In addition, all sorts of extensions and enrichments can be observed making a digital archive a living body and asking for the application of stand-off principles. Another aspect that is special for the humanities is often the necessity to preserve the data, since they represent assets of our cultural heritage and document the state of our minds and societies. This is the reason why in our domain we cannot just speak about repositories or digital libraries but why we need to extend our focus of concern to digital archiving. Since our archives contain sensitive material of various sorts, also

access rights management is necessarily rather complex. For every individual resource access rights need to be definable and a delegation of access rights management needs to be highly granular.

2 Definitions

Before discussing the architecture, usage scenarios, interoperability mechanisms and federation aspects we first should describe a few terms that we will use. We will limit ourselves to the digital domain.

Repository A digital repository is an instance where digital objects of any form can be stored without any predefinition of their utilization and interpretation. In general a repository will maintain a database of descriptive metadata records representing specific characteristics of these objects.

Archive A digital archive is a repository that also has a strategy for the long-term preservation and long-term accessibility of the stored objects and their descriptive metadata standards. Typically adherence to open standards is required to guarantee not only bit-stream preservation, but also interpretability. Major elements of long-term bit stream preservation are technology migration and data distribution.

Digital Library The term "digital library" obviously originates from the library community to describe their extension to the digital domain. Therefore its primary focus is on storing digital publications and making them available via information retrieval functionality. Digital libraries are specialized repositories although the two terms may increasingly overlap.

Descriptive Metadata Metadata is a rather generic term describing any form of added data about other data. Descriptive metadata describes other data with the help of a restricted vocabulary of keywords. This type of metadata description can also be seen as the incarnation of the described object.

Federation of Repositories A federation of repositories is based on a set of agreements or contracts defining trust relations and in case of strong federations even penalties. For the user federations are meant to create an integrated layer of accessibility and responsibility.

When we will use these terms in the following we refer to these pragmatic definitions. Instead of speaking about digital libraries we prefer to speak about repositories in our domain, since the term is less biased. Since the MPI needs to take care of long-term preservation aspects due to its specific resources such as for example the documentation of many languages (currently about 200) that will become extinct in a few years we will use the term "digital archive" in this paper.

3 Background and Architecture

The MPI archive is currently fed by about 100 researchers from the institute, 44 interdisciplinary teams of the endangered languages documentation project DOBES (www.mpi.nl/dobes) and some collaborators from other institutes. Mainly due to the results of a UNESCO study which stated that about 80 % of our recordings about cultures and languages are highly endangered, the MPI decided to give an open archiving service to external researchers as well. Therefore an increasing number of external researchers makes use of the unique facilities and store their data in the MPI archive. Most of these researchers work in small independently operating teams or projects, i.e. the resources stored exhibit a high degree of variation in all dimensions such as the type of resources (texts, media streams with various codecs, time series data with various formats etc), the type of internal structure and the type of relations amongst them. Thus, the archive can be said to be inherently complex.

The key component for maintaining and accessing the archive therefore is the schema-based IMDI metadata set (www.mpi.nl/imdi) which was a result of the broad discussions in the ISLE project (www.ilc.cnr.it/EAGLES/isle/) and which is now subject of standardization in the ISO framework (www.tc37sc4.org/). Due to its archiving role the metadata is primarily stored as XML files, encapsulation in a relational database is only done to support fast searches etc. Every depositor is bound to provide IMDI metadata descriptions with and a canonical tree organization of the resources. The canonical tree organization is being used for various management operations such as copying or curating resources of whole sub-trees and defining areas of responsibilities for setting access rights. Any user can create his/her own linked domain of metadata resources and in doing so create virtual collections of resources from different deposits. The metadata descriptions are open and can therefore be used for browsing, searching and harvesting. Harvesting is supported in two ways: (1) Service providers can harvest the XML files via HTTP or (2) they can use the OAI PMH port (www.openarchives.org/OAI/openarchivesprotocol.html) where also a mapping to Dublin Core is provided. The double function of metadata for the ingest and management operations on the one hand and the user access on the other is indicated in figure 1.

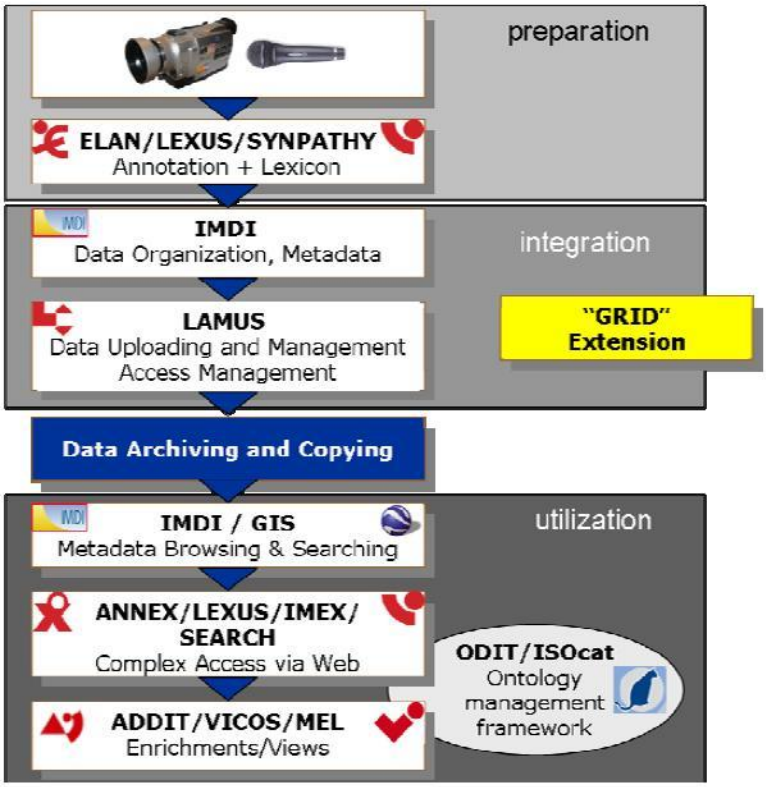

Figure 1. The Language Archiving Technology software suite that also indicates part of the archive's architecture.

A core element for the archive is the LAMUS (www.mpi.nl/lat) component which acts as a kind of gate keeper to create a maximally consistent and coherent archive. LAMUS is used by the depositors to upload resources, resource collections and/or metadata hierarchies. It offers a workspace where people can carry out manipulations before finally requesting to upload the resources or collections. LAMUS will check the file types based on a configuration file where parsers can be associated with the accepted file types for example. It also takes care of generating presentation versions such as compressed streaming formats, offers a component allowing the depositors to set access rights in efficient ways and creates indexes for fast metadata and content search. However, it does not encapsulate any of the uploaded resource, i.e. each resource is available in its format and therefore can be interpreted without additional software. LAMUS also provides means to facilitate data curation which turns out to be very important with respect to quality management. Version handling is a must to deal with resource enrichments.

The LAT software suite will contain increasingly more web applications offering a variety of access options for the different resource types and resources bundles. Singular objects can of course be accessed by navigating in the catalogue down to the

leaves and accessing the objects and visualizing them with the typical browser plug-ins. Accessing media files associated with structured annotations the structure of which is user defined and nevertheless based on a generic schema can be done by ANNEX. ANNEX comes along with a flexible and powerful search tool which makes use of an index that covers all 65 million annotations currently in the archive. For series of images and photos IMEX has been developed to offer efficient access. For lexicons which are all transformed beforehand into the Lexical Markup Format, a flexible ISO standard based on XML, the LEXUS tool can be used. It allows to associate any type of media extension to any lexical entry and attribute and to visualize them.

Currently, components (VICOS, ADDIT) are in development which allow authorized users to make commentaries to any resource fragment and to draw relations between them. VICOS also offers a graphical option to easily navigate in such conceptual spaces. All LAT tools can access the fast indexes via clearly defined APIs.

4 Archiving and Curation

Archiving is based on a clear dedication to open standards such as XML and MPEG, on a storage technology migration every 4 to 5 years and on data distribution. Basic data replication occurs at system level, dynamically in total 7 copies are generated, two residing at the local hierarchical storage management system, 4 at large computer centers distributed across Germany and one copy of the Figure 1 shows the Language Archiving Technology software suite that also indicates part of the archive's architecture. DOBES project data at another institute. This data copying to large computer centers is extended to creating copies at so called regional archives which have been setup all over the world using the LAT software (see figure 2).

Since it is the intention to not use the various repositories as mere backups, but to allow users to make use of the full LAT functionality including extensions and enrichments, a data exchange purely on physical level is not sufficient anymore. LAMUS is currently being extended by a component that can exchange data on logical level which includes operations such as checking checksums, creating versions, updating the unique resource identifier record etc.

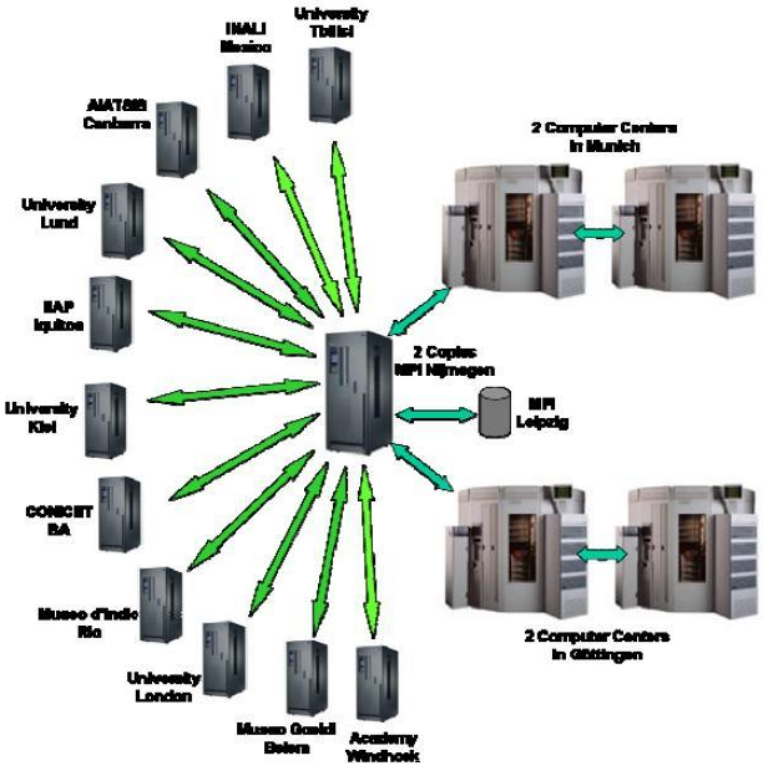

Figure 2. The Current network of archives and regional repositories linked with the archive at the MPI.

The remote archives are synchronizing with the MPI contents in the case of appropriate agreements and the MPI generates backup copies at large computer centers. Currently, all archives are synchronized at physical level which has severe restrictions. A component synchronizing data at logical level is under development which will offer much more options for independent operation of all network partners. Requests to setup seven new regional archives have been received and will be dealt with this year.

Very important is the transformation of all incoming resources into the accepted standard formats. This is a huge and time consuming task, since people often use unconstraint tools to create even structured information. Often also a proper Unicode encoding is not given. In highly dynamic areas such as video encoding the archive has to continuously check appropriate methods. Currently there is a trend towards high definition video typically encoded in H.264 and lossless MJPEG2000 encoding as archive format. Claiming to have an archive requires a strategy for future video encoding that bypasses the well-known concatenation effect that introduces artifacts when transforming between different compressed formats. Thus, only lossless MJPEG2000 encoding will finally satisfy our wishes. However, transforming the

currently 35 Terabytes of the archive to lossless MJPEG2000 encoding and stepping over to high resolution video would require Petabytes of storage capacity. We need to synchronize this transformation with a switch to new cheaper storage media such as holographic disks.

5 Usage and Interoperability

Yet most researchers still work according to the "download first" paradigm which has partly to do with a lack of suitable frameworks to enable a true cyberinfrastructure usage scenario. The various web applications offered by LAT, in particular the archive wide search tools on metadata and content and the enrichment frameworks will change the paradigm slowly, if researchers can rely on their persistence. An increasing amount of explicit API specifications for various components also from other institutes will foster this fundamental change in carrying out research.

Figure 2 shows the current network of archives and regional repositories linked with the archive at the MPI. The remote archives are synchronizing with the MPI contents in the case of appropriate agreements and the MPI generates backup copies at large computer centers. Currently, all archives are synchronized at physical level which has severe restrictions. A component synchronizing data at logical level is under development which will offer much more options for independent operation of all network partners. Requests to setup seven new regional archives have been received and will be dealt with this year. However, building virtual collections by including resources from different deposits introduces a new hurdle that needs to be overcome. Since the metadata format is standardized within the archive no severe structural and semantic interoperability problems will occur. This is different for the content. The archive tries to transform as many resources as possible into generic structural frameworks such as EAF for annotations and LMF for lexicons and by doing so reducing the structural interoperability problems.

With respect to semantic interoperability at the level of tag sets and the encoding of linguistic and anthropological phenomena, other methods need to be developed. In collaboration with ISO TC37/SC4 the LAT team currently is working on the ISOcat data category registry - a web framework to manipulate and access widely agreed linguistic concepts. It is expected that a layered set of such flat registries and shared relation registries will overcome the semantic interoperability problems at the linguistic encoding level. It is obvious, however, that smart tools such as ontology editors to easily create and manipulate "pragmatic ontologies" will be required to facilitate cross-collection operations. Many attempts will be required to tune the current technology components so that a broad group of researchers will accept this new working paradigm. The MPI is working and integrating such components and offering APIs so that others also can make use of them.

6 Archive Federation

In the DAM-LR[1] project the LAT team integrated technology together with its partners that allows users to see one virtually integrated archive based on the distributed holdings. The components that were established were (1) a joint metadata domain based on metadata harvesting and portals, (2) a joint domain of unique and persistent identifiers based on the Handle System, (3) a joint authentication and authorization domain based on Shibboleth and (4) TERENA/TACAR (www.tacar.org/) based certificates to create a domain of trusted servers and services for the various interactions that are required.

This setup allows the user to create virtual collections in a seamless way since he/she only needs to authenticate him/herself once at his home institute. Secure exchange of user credentials will take care that all archives can perform authorization checks based on the transmitted attributes. Establishing a real federation would require establishing a domain of mutual trust based on agreements. The nature of such agreements was studied, but making contracts must be left to persistent research infrastructures.

7 Research Infrastructures

Consequently, the LAT team was an active partner from the beginning to establish the persistent CLARIN (www.clarin.eu) infrastructure. The investments to come to an integrated and interoperable domain of language resources and technology components are so high that it is necessary to establish an organizational framework that is persistent. The CLARIN initiative emerged as one of the 35 ESFRI (http://cordis.europa.eu/esfri/) proposals that were accepted by the EC. It currently covers 112 institutions from 35 European countries and first discussions about national long-term roadmap towards eHumanities have already been started in some countries. Stable centers with long-term existence guarantees offering resources and services will form the pillars of such an infrastructure.

[1] Distributed Access Management for Language Resources was an EC funded project from 2005 to 2007 where 4 linguistic archives where coupled by typical federation technology components. In this successful test it was not the intention to create formal agreements to make it a real federation. This was left to the infrastructure project CLARIN.