short Guide

Common Language Resources and Technology Infrastructure

February 2010

Relevance

CLARIN community

for all communities

REPLIX

What is it?

REPLIX is a project studying and implementing the next level in grid based safe data replication and synchronization. This is important for preserving our petabytes of language resources for the future. REPLIX will focus on issues such as data authenticity, data integrity, maintaining access rights and data organization since they are so crucial for proper data management in a distributed environment. Replication in this context means copying data in one direction thus realizing a master-slave relationship, while synchronization is much more challenging in a world where copies of collections are extended and new versions are created at various places. To handle replication and synchronization in a proper model for digital object (DO) identity, REPLIX will make use of information associated with persistent identifiers (PID).

What is it for?

Users increasingly often want to hand over data management to repositories that can offer persistent access services, thus relying on a proper treatment of their resources based on trust declarations.

They want to

- Store their data in safe and persistent environments which means storing data at various locations and in a way that authenticity and integrity is guaranteed.
- Work on a local repository and synchronize with a distributed data infrastructure when the local work is finished.
- Be ensured that their complex metadata structure is preserved in the new context of the distributed data infrastructure.
- Be ensured that the PID records associated to their data are kept up-to-date pointing to the copies of the objects and storing characteristic data about the object.

Who can use it?

The REPLIX project will implement and test such a replication and synchronization infrastructure so that we can study the underlying design principles and software components. As such it primarily addresses the needs of repositories and archives which are to be trusted with managing a depositor's data. REPLIX assumes that all data centers have their own repository system and the desire to keep it. Following from this assumption the REPLIX replication and synchronization layer needs to be flexible enough to couple to any existing repository system independent of the community that provide persistent identifiers and metadata descriptions to structure their data collections.

When can it be used?

Currently four pilot tasks have been planned for 2010 by making use of the MPI archive as an example to create insight in the basic challenges in this domain, such as federated synchronization, security and trust, synchronization of the complex metadata, etc. Based on these pilot tasks where we are using iRODS to check its functionality and robustness we expect a clear basis for design and implementation decisions in the autumn 2010 and to deliver fi rst proof of concept systems in 2011..

How does it work?

REPLIX aims at creating solutions which introduce a thin, non invasive layer for the data replication and synchronization task between collection structures residing in different repositories. This layer should be able to access the repository and data grid information based on proper programming interfaces (APIs). In general REPLIX aims to create a general solution, not specific to a single context. But all functionality will be tested with the MPI language archive as a concrete example with a wide variety of data types ranging from media and time series files to textual annotations grouped in sessions and recursively defined collections (see figure 1). A user should be able to select a node in such a hierarchical structure representing a certain collection and give the command to replicate this collection to a location in the hierarchical structure of a destination archive, probably located at a different physical location (see red arrow). REPLIX needs to take all measures to carry out such a copying at a logical level. This includes copying the physical fi les to the correct destination, making sure the hierarchical structure at the destination is properly updated, making sure the copy is authentic, making sure the access rights are transferred to and applied at the destination, the associated PID records are up to date, etc. Regular audits will compare the checksums of the stored objects against the known values which are associated with the PIDs.

Each repository will have developed layers of typical upload, management and access tools that interact with the archive such as the Language Archiving Technology (LAT) at MPI. These software layers are represented by one abstract layer in the diagram. For REPLIX it is essential that its functionality may not intervene with all the software layers that have been developed so far to ensure the proper operation of the repositories.

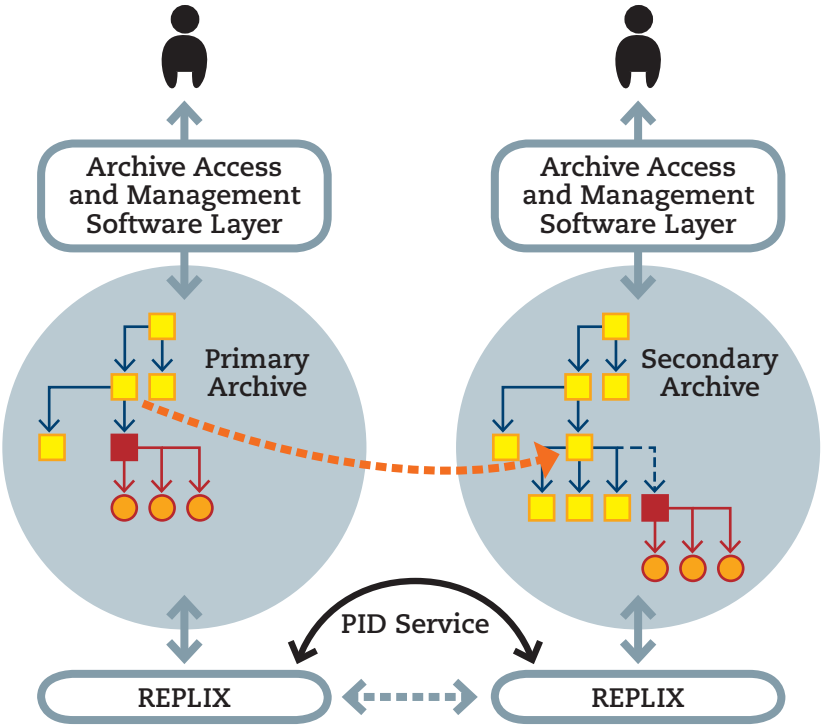

Figure 1: REPLIX Process

Collection
Session
Media file,
Time series file,
Textual annotation

Who is responsible?

Both Max Planck for Psycholinguistics, representing CLARIN and DOBES, and RechenZentrum Garging, representing DEISA, are collaborating in this project. REPLIX will collaborate with other interested partners and also with industry.

Whom to contact?

For all aspects the following addresses can be contacted:

Willem Elbers (MPI): willem.elbers@mpi.nl

John Alan Kennedy (RZG): jkennedy@rzg.mpg.de

Where to find more information?

Further information about the REPLIX project is available via the project's website:

REPLIX: http://www.mpi.nl/replix

MPI Archive: http://www.mpi.nl/resources/data

LAT: http://www.mpi.nl/resources/tools

Responsible for the content:

Willem Elbers
MPI for Psycholinguistics
Wundtlaan 2, 6525 XD Nijmegen, NL
Website: http://www.mpi.nl/replix
Email: willem.elbers@mpi.nl