# ECP-2008-DILI-518001

## BHL-Europe

## Technical Note

# Proof of concept LOCKSS (1.Analysis)

| | |
|---|---|
| **Deliverable number** | *TN-SPRINT1-106* |
| **Dissemination level** | *Confidential* |
| **Delivery date** | *2010-06-30* |
| **Status** | *Draft* |
| **Author(s)** | *Walter Koch* |



## *e*Content*plus*

---

## *Table of contents*

# 1 Document History

## 1.1 Contributors

A discussion about this issue was initiated and the following persons provided input that was used for the present document.

| Person | Partner |
|---|---|
| Walter Koch | AIT |

## 1.2 Revision History

| Revision Date | Author | Version | Change Reference & Summary |
|---|---|---|---|
| 2010-07-15 | Walter Koch | 0.1 | 1. Draft |
| | | | |
| | | | |

## 1.3 Distribution

This document has been distributed to:

| Group | Date of issue | Version |
|---|---|---|
| BHLE – technical group | 2010-07-16 | 0.1 |
| | | |

## 2  Purpose of this document

This document describes the first findings about a possible integration of the LOCKSS-System (Lots of Copies Keep Stuff Safe) into the BHLE System. It should provide a basis for discussion whether it is useful to look deeper into this issue or stop the investigation in this direction in favour of another solution, eg. Bittorrent.

## 3  Background and Context

### 3.1  The BHLE-Requirement

The Global BHL System is based on the replication of the large amount of static data (whole BHL data sets) in different regions (America/US, Europe/EU, Australia, Asia/China). There are different methods to achieve this, eg using Bittorrent or LOCKSS-Boxes.

The high level architectural design as discussed in several technical meeting looks like this: (https://bhl.wikispaces.com/BHL-E_WP3_ArchitectureDiagrams - visited: 2010-07-15) The replication takes place in the middle light blue box (green cycle "**B**") in the digramme below. This component could be considered as being the aggregation platform which contains data from all regions. Data belonging only to a region is included in the light blue block at the bottom (grey cycle "**P**").
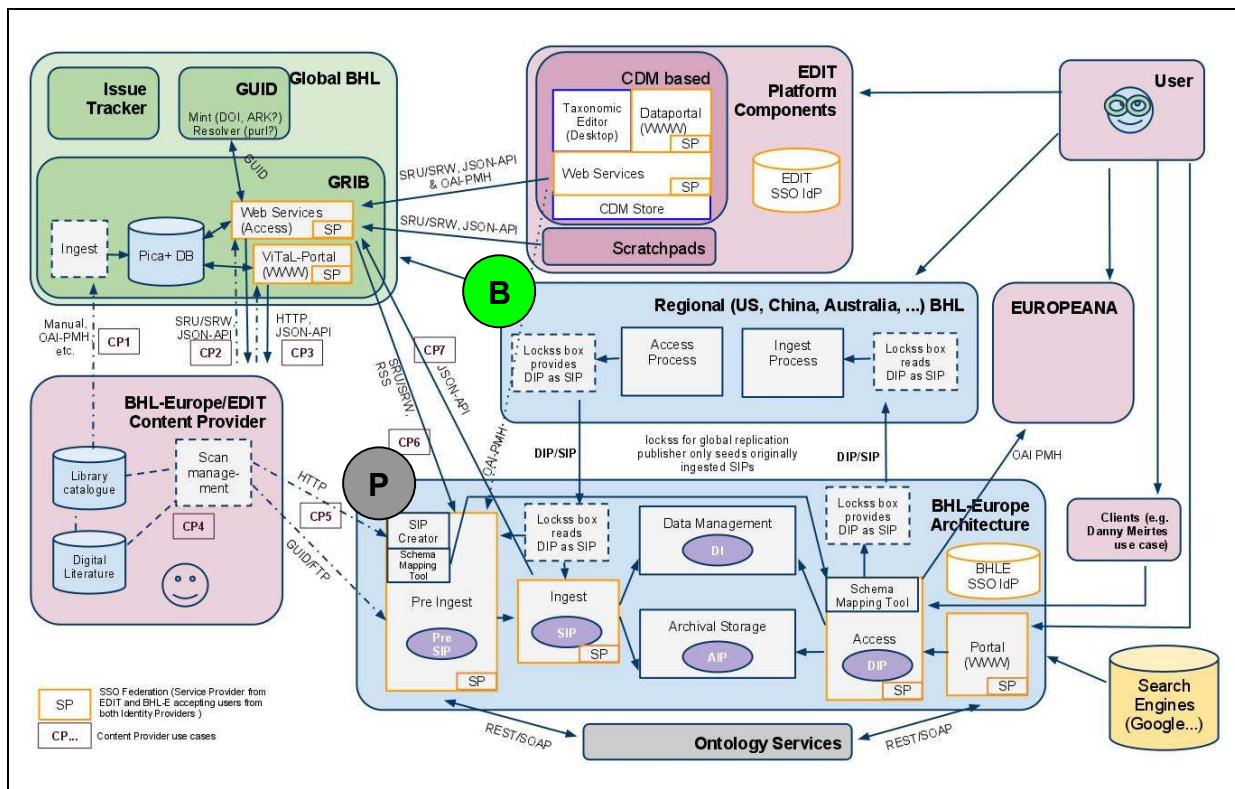
### *3.2   LOCKSS - Lots of Copies Keep Stuff Safe*

In the library world for several years a system for Distributed Digital Preservation has been successfully developed, the LOCKSS System. The system *"[…] provides an OAIS-compliant, open source, peer-to-peer, decentralized digital preservation infrastructure. It is format-agnostic, preserving all formats and genres of web-published content, provided the content has an authoritative version."* (http://lockss.org/lockss/How_It_Works#Providing_Access ; URL visited: 2010-07-15).

In an earlier publication of  Vicky Reich (2001) the Data Flow in a LOCKSS based distributed system looks like this (figure 1):
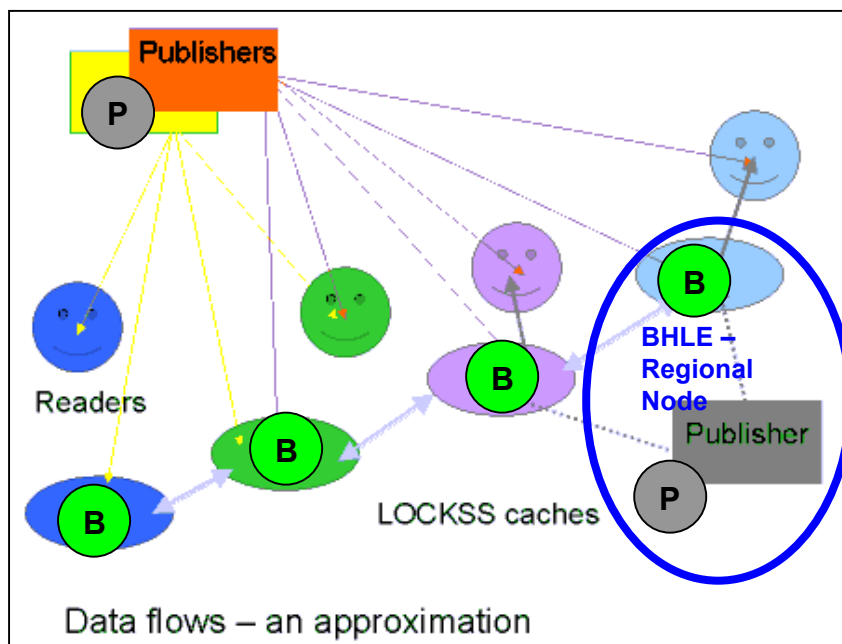


*"Figure 1: In this example, each LOCKSS cache (oval) collects journal content from the publisher's web site as it is published. Readers (circles) can get content from the publisher site. When the publisher's web site is not available (gray) to a local community, readers from that community get content from their local institution's cache. The caches "talk" to each other to maintain the content's integrity over time."*

( http://www.dlib.org/dlib/june01/reich/06reich.html ; URL visited: 2010-07-15).

## 4   Proposed approach

Locking into the concepts behind LOCKSS and possible implication to the BHLE Project is only one option but will be further pursued in this document. All statements are made on a first conceptual basis and have to be discussed in the appropriate BHLE groupings.

### *4.1   LOCKSS and the BHL - Architecture*

Looking into paragraphs 3.1 and 3.2 it is assumed that on the "regional level" BHL-Europe provides a regional (European) platform which aggregates content from European countries and contributes this content to the global BHL System. This regional system will furtheron called a "BHL-node".  A (the European) BHL-node consists of the two components labeled in par. 3.1 with "P" (to be delivered by the EU-funded project) and "B" respectively. This node interacts with the contributing archives (BHLE-content providers) as well as with "central components" like an ontology node, a "deduplication system" (GRIB), etc.

Aligning this situation with the LOCKSS concept the BHLE-component "P" could be considered as a "publisher" in the LOCKSS environment and the component covering the global BHL data set as a "LOCKSS-Box" containing the cached data from all other Boxes. A "publisher-component" in this concept is responsible for all data generated on a regional level and to be subscribed by the BHL-LOCKSS-Boxes and may be connected to a local production and control system (eg: MetaArchive, PeDALS, or similar implementations of a "PLN" – Private LOCKSS Network).

## 5   Organisational and Technical Considerations

Implementing a Private LOCKSS network one has to look into organisational and technical requirements and constraints.

### 5.1   Technical Issues

The whole LOCKSS Software is available in the open source, the software licence is:

The software can be downloaded from:  http://sourceforge.net/projects/lockss/ (URL visited: 2010-07-16. Note: the CD image was not the actual one used at that time). The source code is available from: http://lockss.cvs.sourceforge.net/lockss/ and contains four components:

- The cache manager
- The lockss daemon
- The lockss platform
- The lockss-ui

The whole software was downloaded as GNU tarball and installed on virtual Ubuntu 10.04 box. There have been only minor problems compiling the whole stuff in this environment.

### 5.1.1  The LOCKSS-Box

The ISO-image which includes the platform and daemon has been downloaded from http://www.lockss.org/release/lockssCD280.iso (URL visited: 2010-07-16) and installed as virtual image on a Ubuntu-host machine. The installation needs no special expertise, is straight forward and can be done within half an hour. A test box is now available and running at:  http://lockss.ait.co.at:8081/  (user/password:  lockss/lockss).

### 5.2  Organisational Issues

There are some requirements for running a PLN:

- Members of a PLN have to be member of the LOCKSS-Alliance (annual fee up to 10.800 US$ per member, depends on organisation type)

> **Support for PLN members**
>
> All members of a PLN must be LOCKSS Alliance members and group discount rates are available. Support includes consulting on hardware and software for LOCKSS boxes, site design and plugin design, implementation and testing, server hosting of configuration parameters, plugins, and title database, etc., and assistance with proxy integration or content export.

- There must be a minimum number of partners (six) running a box in the network.

# 6   Proposed Actions

Before decision is taken to go ahead with this option (running a distributed BHL archive on global level)  some points have to be clarified:

## 6.1   Clarifications needed

1. Can the software be used without being member of the LOCKSS alliance (in case there is no support needed) ?
2. If answer iy Yes – is it possible to get support via the German LuKII project (which has to be agreed by the LOCKSS directors)
3. Can the software be used without any modifications (eg using a simple SIP like defined for bagit)
4. Can a co-operation agreement be made with other PLN providers( eg MetaArchive).

## 6.2   Actions

1. Decide whether it is usefull to go ahead with the LOCKSS approach (Technical Meeeting in London)
2. If Yes -> Clarify the points outlined in 6.1 (first contact with Prof.Michael Seadle from the LuKII project has already been established). If No -> stop.
3. Setup a working group preparing the implementation of a PLN-like distributed archival network and integration into the overall BHL architecture.