PIDs for CLARIN

Daan Broeder
CLARIN / Max-Planck Institute for Psycholinguistics

CLARIN D Tutorial Sept. 2011

Contents

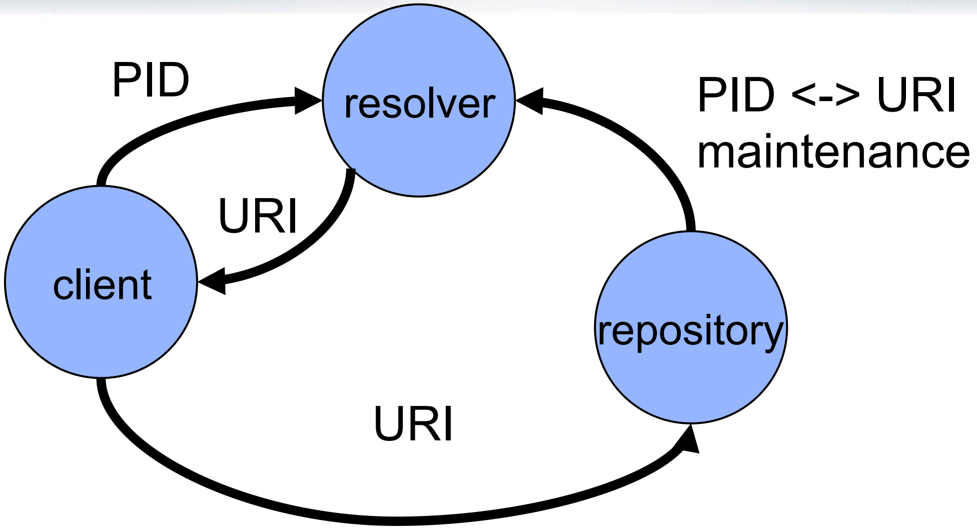
- Persistent Identifiers
- CLARIN requirements & policy
- PIDs & Granularity
- PIDs & Versioning
- How about the user?

Referencing Resources

Standard we use URIs for referencing resources. However: the resource gets moved – domain name change or file system changes you get a ‘404 not found’ (link rot)

- Problem for embedded references inside the repository system/archive
- …but especially outside the archive
- Can be seen as an organizational problem
- But difficult to solve, hence the PID frameworks

Referencing Resources



- Give every resource a unique persistent identifier: PID
- Every PID associated with one (or more) URLs
- Resolving process built into applications or available through plug-ins.

This comes at a cost:

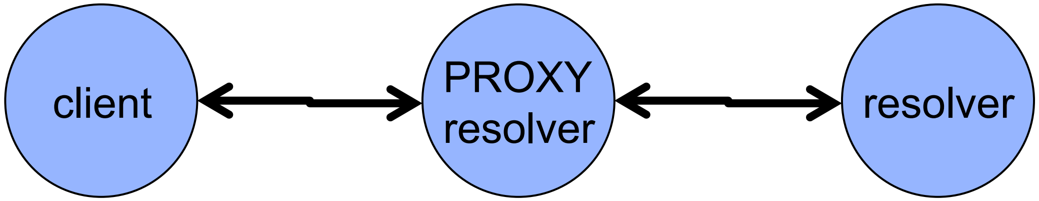
- Added layer of infrastructure
- Must be managed
- Must run with high availability
- Must be very sure that this can be handled by our archives also in the long term.
- But can be used for extra services

Actionable PIDs

- PIDs are not automatically *actionable* (clickable) without special browser or document reader plug-ins.
- PIDs can be made actionable by *URLifying* them: embedding them in a HTTP URI syntax that redirects the application to a special HTTP proxy resolver

MYPID -> http://HTTP-PROXY-RESOLVER/MYPID

- The proxy resolver uses the PID resolver info to again redirect the application to the URI of the resource



Cool URIs

- W3C advice: use ‘Cool URIs’
  - Persistent identification is a organizational problem
  - Your domain name should be stable
  - Your resource name should be well chosen
  - Use for instance apache’s mod_rewrite to map URIs on changing file paths.
  - Think of branding, have your resources start with **http://yourarchive.nl/…**
  - URIs are always ‘actionable’
- However:
  - not all of us have such a stable domain name
  - Using mod_rewrite your config file can get very big
  - … and there can be other advantages using a PID framework

PID Frameworks

- PURL – Persistent URL
  - mid 1990ties by librarians (OCLC)
  - Uses simple HTTP redirect to refer HTTP clients from a central PURL resolver to ‘changing’ resource locations
  - *http://purl.org/dc/elements/1.1/*
- HS – Handle System
  - 1994 CNRI (Bob Kahn ao.)
  - Independent of HTTP
  - Robust distributed resolver system
  - Possibly link multiple URIs and other information to a handle
  - Offers mechanism, does not impose policy
  - *1839/00-0000-0000-000F-56C9-1*
  - Used by LC, DOI, EPIC, DataCite

PID Frameworks

- URN
  - Widely accepted standard for unique identification
  - But no widely accepted global resolver, but see PERSID project
- XRI
  - Proposed by OASIS 2005
  - Encourages semantics in the PID
  - Useable for all types of identifiers: email, persons etc.
  - Only syntax spec, no resolver (part. impl. probably exist)
  - W3C intervened preventing the XRI 2.0 standard adoption, now only urlified XRIs are proposed.
  - *xri://broadview.library.example.com/(urn:isbn:0-395-36341-1)*
  - *xri://@Jones.and.Company/(+phone.number)*

PID Frameworks: ARK

Archival Resource Key *John Kunze, CDL 2003 -*

- *[http://NMAH/]ark:/NAAN/Name[Qualifier]*
- An ARK identifier is associated with three services:
  - Providing a link to the resource;
  - Providing a link to the resource's metadata; and
  - Providing a link the resource provider's promise about its persistence;
- Its naming scheme is constrained to discourage semantics in the identifier;
- It accommodates resource part identifiers and possible different resource representations;
- ***Status of global ARK resolver is unknown to me***
- ***Kunze has have a new project: EZID that does both ARK and DOI***

PIDs for CLARIN

From the Short Guide “criteria for CLARIN centers”

- Centers need to associate persistent identifier with their resources that can be resolved and that can be used to test authenticity
- Guarantee to deliver the same content for the same identifier *(This is preferable but perhaps not tenable)*
- Explicit statements about the duration and the quality of the services
- In general CLARIN will expect that if a center stops offering services to CLARIN these service will be transferred to another center to guarantee continuity for the user community

PID framework selection

	Standard	Robust Software	Resolution System	Resolution Type	Security Admin	Assoc Info	Cost
URL	RFC2616	no	yes (DNS)	single	no	no	no
URN:ISSN	ISO2397	no	no	?	no	no	no
URN:ISBN	ISO2108	no	no	?	no	no	no
URN:NBN	RFC3188	no	no	?	no	no	?
PURL	no	no	yes	single	no	no	no
Handle	RFC3650	yes	yes	multiple	yes	yes	little
DOI	Z39.84…	yes	yes (Handle)	multiple	yes	yes	large
ARK	no	no	(yes)	multiple	(no)	yes	?
info URI	RFC3668	no	no	?	no	no	no
XRI	yes	no	no	?	no	?	?

CLARIN PID requirements

- Attach multiple URLs to a PID
- Allow part identifiers for constituent objects. Granularity issue.
- Allow attaching of extra data records to the PID (MD5 check,…)
- Actionable (URLified) PIDs
- HTTP proxy for resolving (use port 80 only)
- REST or SOAP interface for administration of PIDs from applications
- Secure administration
- Delegation of PID administration to other organizations

- Distributed, robust, highly-available, scalable
- No single-point of failure
- Acceptable non-commercial business model
- Control by user community

Choice for HS

- MPI PL, MPG, CLARIN, … made a choice for the Handle System as the basis for a PID service.
- Has been a process of years, organizing seminars, making evaluations, talk to competing frameworks, argue with W3C, MPG IT council discussions, having hands-on experience.
- Compatible ISO 24618 PISA: Persistent Identification Sustainable Access for Language Resources

However other opinions exist, therefore:

- NO ONE is obliged to use the HS
- but CLARIN centres need to be sure that
  - what you do is robust and persistent
  - it can fulfil the essential requirements
  - it can resolve your PID

- Why not register two PIDs ?

Handle System

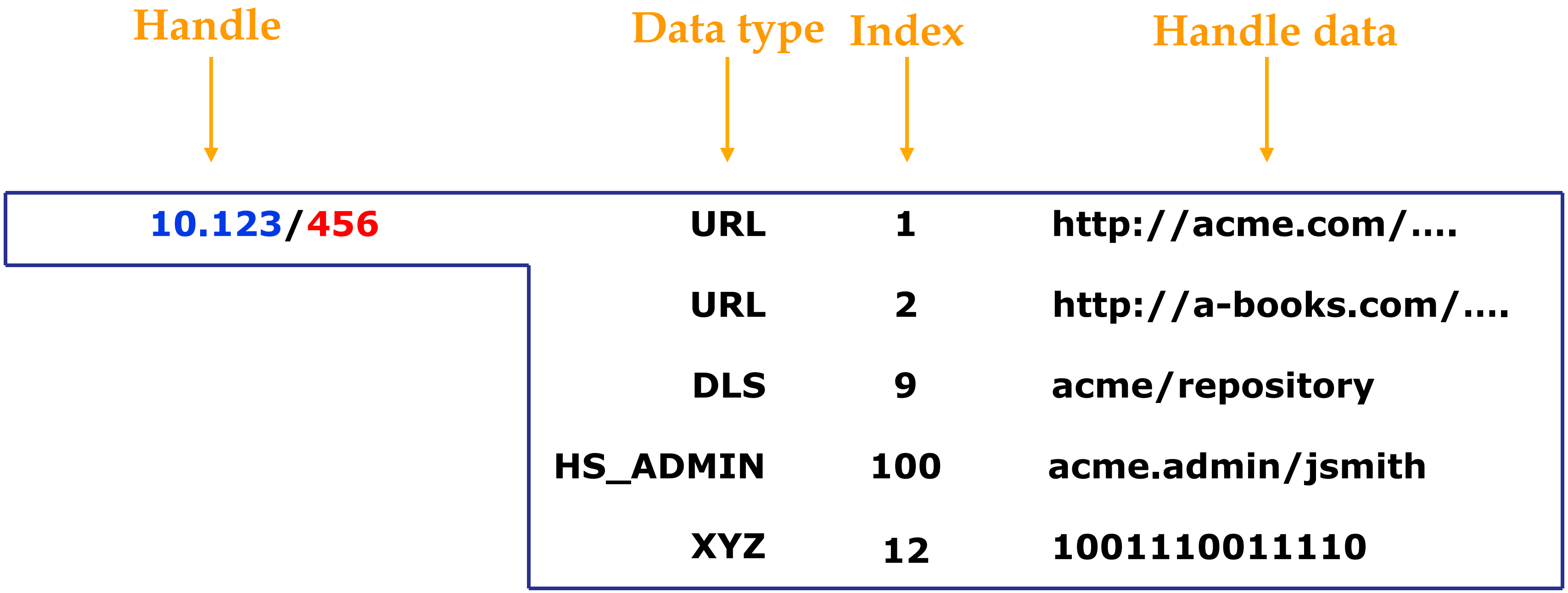
CLARIN proposes Handle System from CNRI as PID framework but not impose it!

- Resolving service is distributed, scalable, secure, optimized for speed.
- Enables association of one or more typed values, e.g., URL, with each PID/Handle
- Supports associating more than one URL per PID
- Easy transfer of management responsibility if needed!
- Also used by the publishers in DOI but is independent

State of affairs:

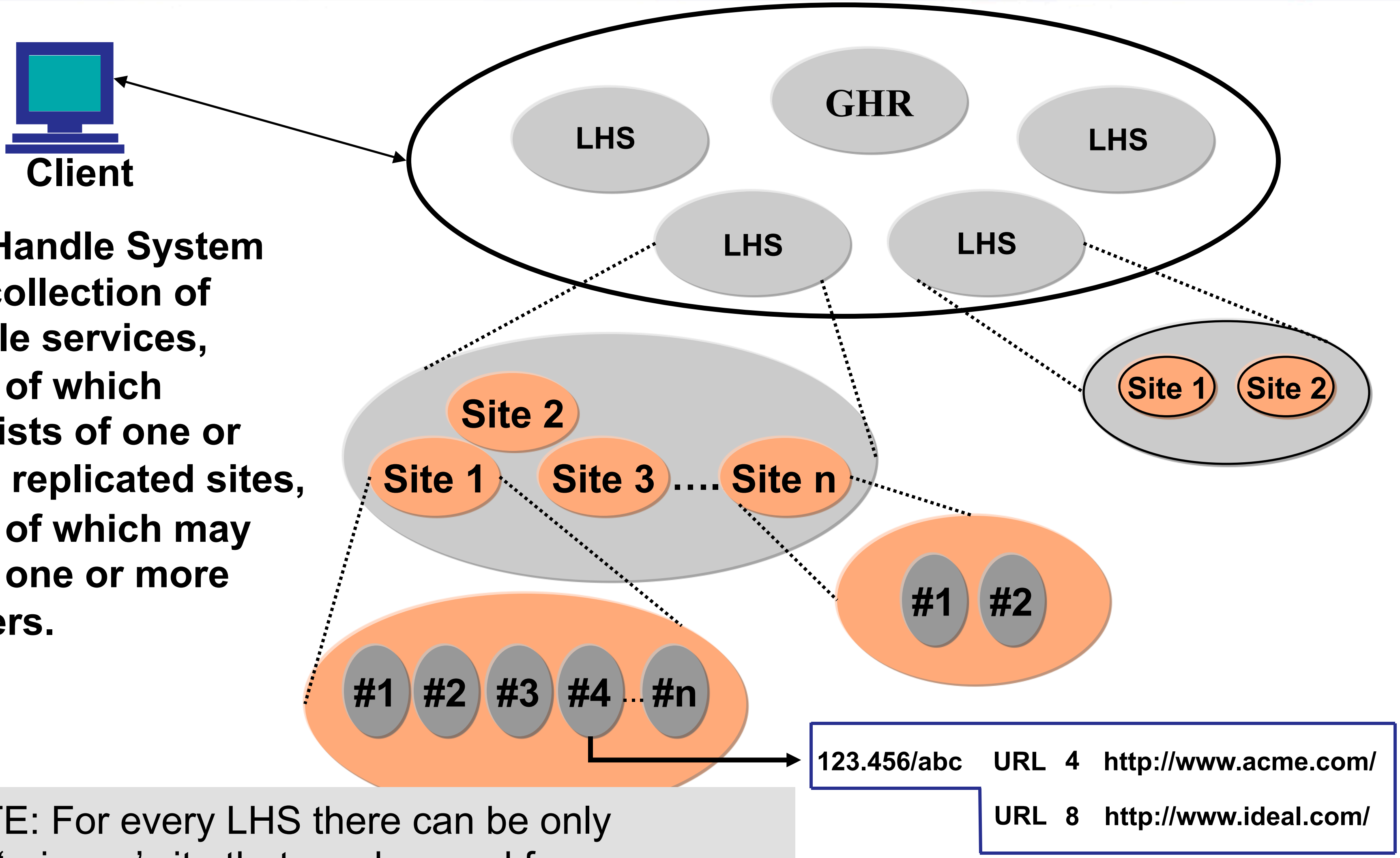
- MPG (GWDG) will run a EU GHR mirror; ‘political’ independence
- CLARIN supports EPIC that will offer PID/handle service to MPG, CLARIN, DARIAH centers (+ other European Science Projects)

Handles Resolve to Typed Data



Handle	Data type	Index	Handle data
10.123/456	URL	1	http://acme.com/….
	URL	2	http://a-books.com/….
	DLS	9	acme/repository
	HS_ADMIN	100	acme.admin/jsmith
	XYZ	12	1001110011110

Handle Resolution



123.456/abc	URL	4	http://www.acme.com/
	URL	8	http://www.ideal.com/

The Handle System is a collection of handle services, each of which consists of one or more replicated sites, each of which may have one or more servers.

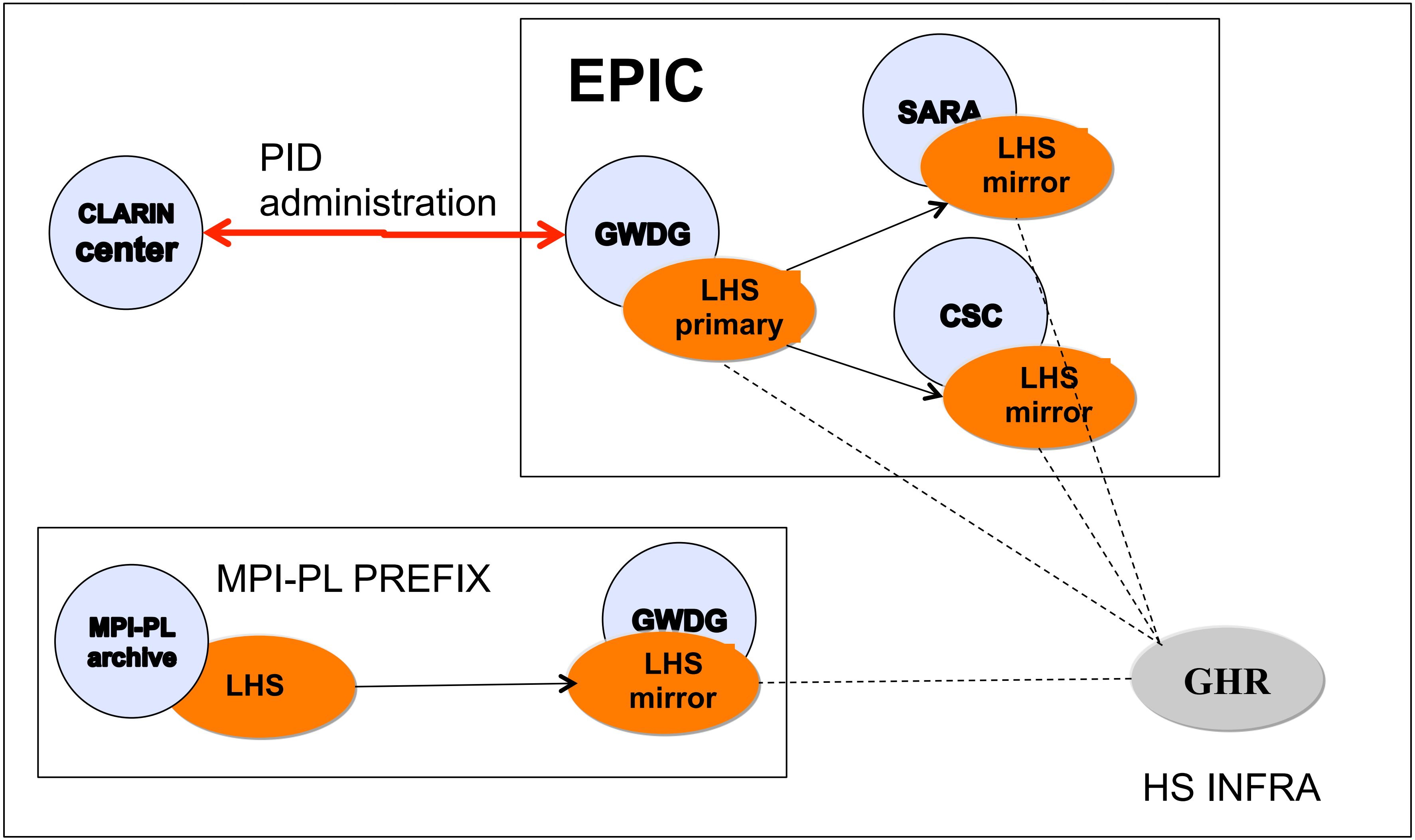
NOTE: For every LHS there can be only one ‘primary’ site that can be used for administration of handles. The others are mirrors.

Central shared PID service

- Possible for every repository to run its own PID service
- Not every organization is willing or able to that
- Increased reliability by replicating services
- etc..

- Start of 2009 the GWDG started a PID service for the MPG institutes (also accessible by other EU scientific projects) based on the HS
- This was absorbed in the EPIC consortium: SARA, CSC
- New members are joining DKRZ, RZG, STFC, …

PID service


EPIC
SARA
LHS mirror
PID administration
CLARIN center
GWDG
LHS primary
CSC
LHS mirror
MPI-PL PREFIX
MPI-PL archive
LHS
GWDG
LHS mirror
GHR
HS INFRA

Some issues

- ‘Political’ independence
  - EU GHR & HTTP proxy (no single point of failure)
  - HS policy board; this now being arranged via ITU
  - Patent ownership, domain ownership; all will be resolved
- Acceptance of HS by W3C
  - Forget that, look at XRI case☹
- Support for part identifiers; available in latest HS version
- Support for multiple administrative sites
  - Is being studied, a multiple primary will be made possible

CLARIN PID records

What handle records does CLARIN need?

Lets be conservative and not use our fantasy

- Multiple URIs
- MD5 checksum to check authenticity
- Pointer to a metadata record
- Persistency promise (or do that in the metadata)

How many PIDs do we need?

How many PIDs can we manage?

GRANULARITY

Part identifiers

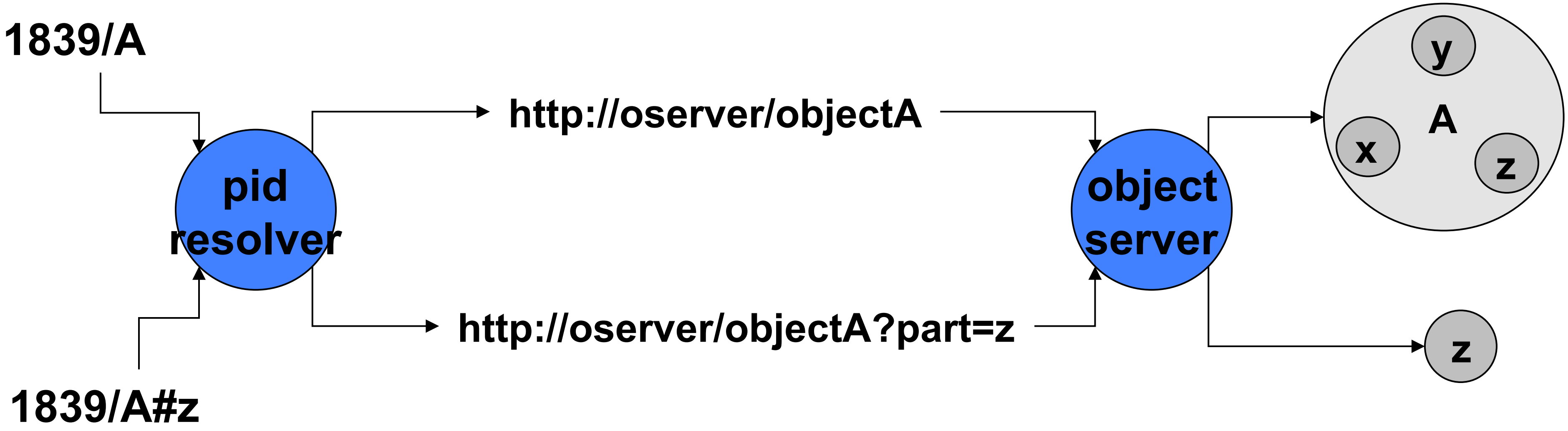
Feature is called handle templates available in latest HS version

ALTHOUGH I USE THE # CHARACTER, THESE ARE NOT URI FRAGMENTS

o issue a pid for each part (think of
es in a lexicon). So use part

an make an adequate translation
objectA?part=z" This requires
xibility from the resolver to
ate the object server.

- The syntax of "Z" should be standard for the specific data type. Loan from existing fragment identifier syntax standards.



HS & Fragments examples

http://corpus1.mpi.nl/fragments.html

Handle examples

- **Annex 1839/URLWARP1 [Data]**
- **ISOcat 1839/URLWARP2 [Data]**
- **Handle fragment examples**
- **Annex 1839/URLWARP1@start=55,length=5 [Data]**
- **Isocat 1839/URLWARP2@lang=fr [Data]**
- **Manual suffix examples**
- **Annex 1839/URLWARP1@time=55000&duration=5000 [Data]**
- **Isocat 1839/URLWARP2@objectLanguage=fr [Data]**

Granularity Recommendations I

The following recommendations are designed to encourage efficiency and promote interoperability with other naming schemes (from ISO 24618)

- If there is an existing identifier scheme for a type of resources, for instance, ISBN, this level of granularity should be retained, which is to say that no new PIDs should be issued without very good reasons, such as for chapters. Chapters would preferably be addressed using part identifiers in conjunction with the PID of the book.
- If the resource is associated with the complete content of a digital file, an individual PID should probably be assigned for this resource.

Granularity Recommendations II

- If the resource is autonomous and exists outside a larger context, an individual PID should probably be assigned for this resource.
- *If a resource should be citable apart from any containing resource, an individual PID should probably be assigned for this resource.*

These recommendations are, however, subject to the needs of resource creators with respect to the level of granularity they deem suitable to the specific resource environment.

Should there be any relation?

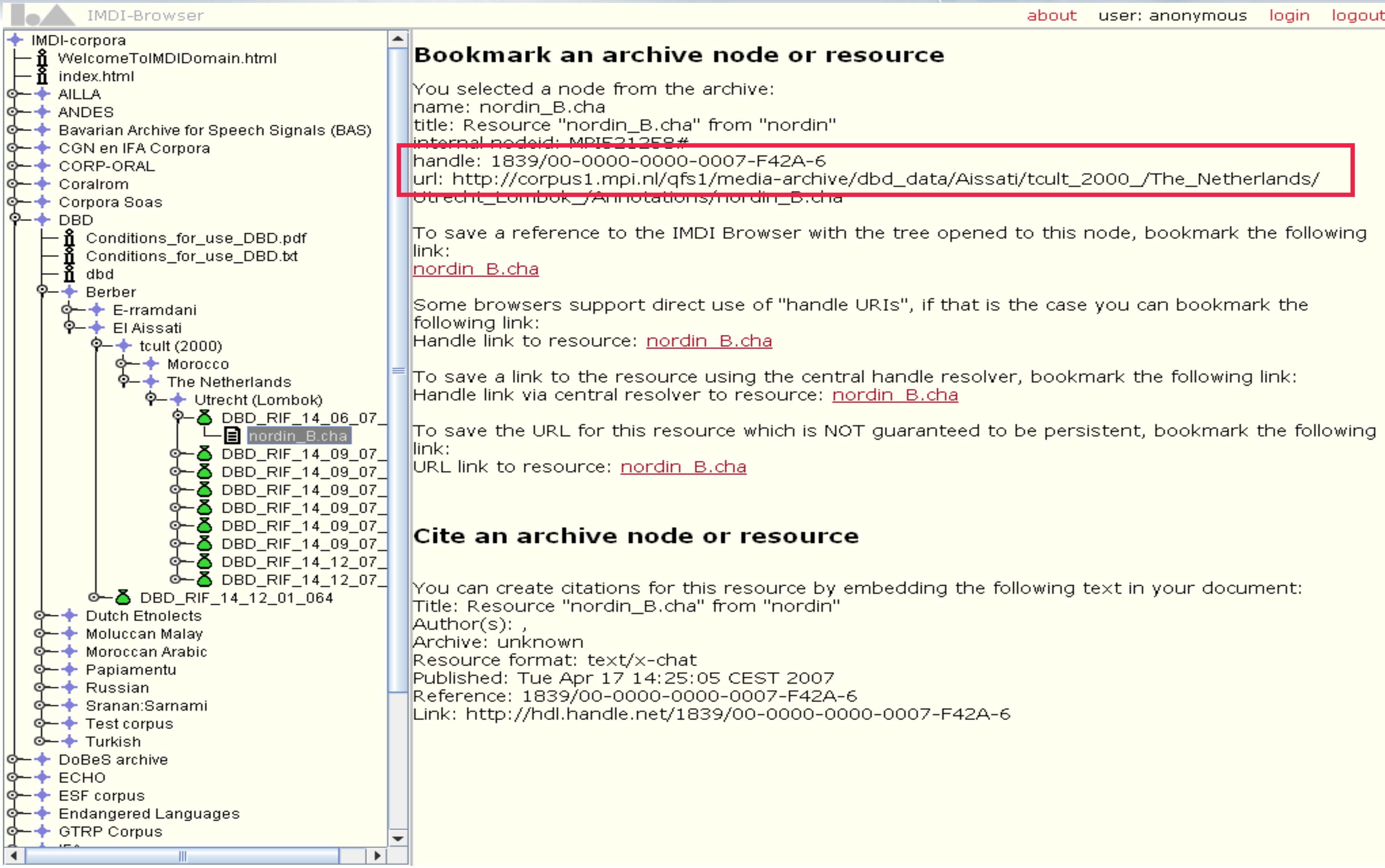
PID & VERSIONING

Versioning

Some ideas:

- Have a separate PID for every version, by default users should be served the same version they bookmarked.
- Questionable to mix versioning with the PID infrastructure
  - Put it in the PID syntax? Myprefix/myobject.V1 (like ARK)
  - Put information in the handle record referring to older/newer versions?
- Make a separate service to page through subsequent versions
- Pointers to an older version can be stored in the newer version metadata

What does the user see?



Bookmark an archive node or resource

You selected a node from the archive:
name: nordin_B.cha
title: Resource "nordin_B.cha" from "nordin"
internal nodeid: MPI521258#
handle: 1839/00-0000-0000-0007-F42A-6
url: http://corpus1.mpi.nl/qfs1/media-archive/dbd_data/Aissati/tcult_2000_/The_Netherlands/Utrecht_Lombok_/Annotations/nordin_B.cha

To save a reference to the IMDI Browser with the tree opened to this node, bookmark the following link:
nordin_B.cha

Some browsers support direct use of "handle URIs", if that is the case you can bookmark the following link:
Handle link to resource: nordin_B.cha

To save a link to the resource using the central handle resolver, bookmark the following link:
Handle link via central resolver to resource: nordin_B.cha

To save the URL for this resource which is NOT guaranteed to be persistent, bookmark the following link:
URL link to resource: nordin_B.cha

Cite an archive node or resource

You can create citations for this resource by embedding the following text in your document:
Title: Resource "nordin_B.cha" from "nordin"
Author(s): ,
Archive: unknown
Resource format: text/x-chat
Published: Tue Apr 17 14:25:05 CEST 2007
Reference: 1839/00-0000-0000-0007-F42A-6
Link: http://hdl.handle.net/1839/00-0000-0000-0007-F42A-6

In IMDI world 1 Mio PIDs

```
<?xml version="1.0" encoding="UTF-8"?>
<METATRANSCRIPT ArchiveHandle="hdl:1839/00-0000-0000-0005-82B0-2"
    Date="2006-07-18" FormatId="IMDI 3.0"
    Originator="Editor - Profile:SESSION.Profile.xml" Type="SESSION" Version="1"
     xmlns="http://www.mpi.nl/IMDI/Schema/IMDI" xmlns:xsi="http://www.w3.org/2001/
     XMLSchema-instance" xsi:schemaLocation="http://www.mpi.nl/IMDI/Schema/IMDI ./
     IMDI_3.0.xsd">
 <Session>
       <Name>DBD_RIF_14_12_01_064</Name>
       <Title>Dutch Bilingualism Database, Ethnic Dutch, Session 64</Title>
     ……….

<MediaFile>
<ResourceLink ArchiveHandle="hdl:1839/00-0000-0000-0004-DC6B-0"> http://
     corpus1.mpi.nl/qfs1/media-archive/dbd_data/boumans/T-Cult/Metadata/../Media/
     dbd_rif_14_12_01_064.wav</ResourceLink>
     ……….
```

Thank you for your attention

CLARIN has received funding from
the European Community's Seventh Framework Programme
under grant agreement n° 212230