The CLARINO metadata infrastructure to be – issues and considerations

5. June 2012

Oddrun Pauline Ohren

National Library of Norway

Nasjonalbiblioteket

Content

Background

Our task (WP2) - in context of the overall CLARINO service portfolio

What do we need?

- Some requirements to the metadata infrastructure

Alternative approaches and solutions

How to proceed from here

Background

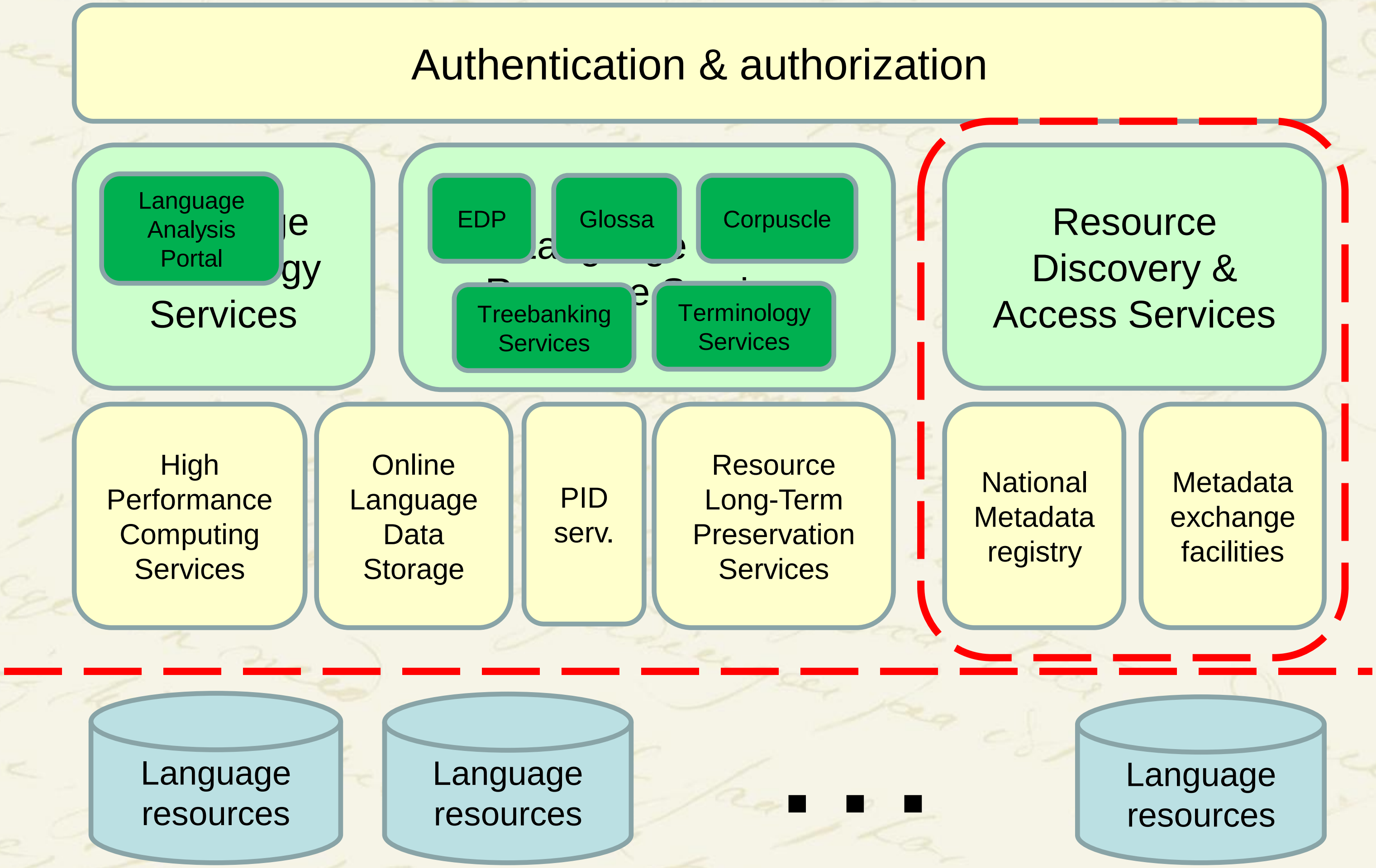
- The Language Technology Resource Collection for Norwegian (Språkbanken)
- NB was mandated by the gvt in 2010 to
  - Collect/create resources, including descriptive metadata
- Work in progress
  - No metadata registry yet
  - Some resources registered in the META-NORD initiative
- Språkbanken’s resources are to be included in CLARINO
  - but also to be regarded as an LR collection in its own right

Our task

“NB will set up a national data registry facility providing *DC mapping and an OAI PMH gateway for metadata exchange with other national CLARIN registries and will set up OAI PMH or XML based harvesting from all Norwegian nodes providing resources. NB will set up an interface allowing users to search the registry for language resources based on specified metadata. NB will set up a national long-term archive coupled to the registry, with secure redundant storage and media migration. The registry and archive will be operational by the end of Y1, but interaction with other components will be implemented as needed in subsequent years.”*

(From the CLARINO DOW)

CLARINO services overview



Authentication & authorization

Language Analysis Portal

ge
gy
Services

EDP

Glossa

Corpuscle

Treebanking Services

Terminology Services

Resource Discovery & Access Services

High Performance Computing Services

Online Language Data Storage

PID serv.

Resource Long-Term Preservation Services

National Metadata registry

Metadata exchange facilities

Language resources

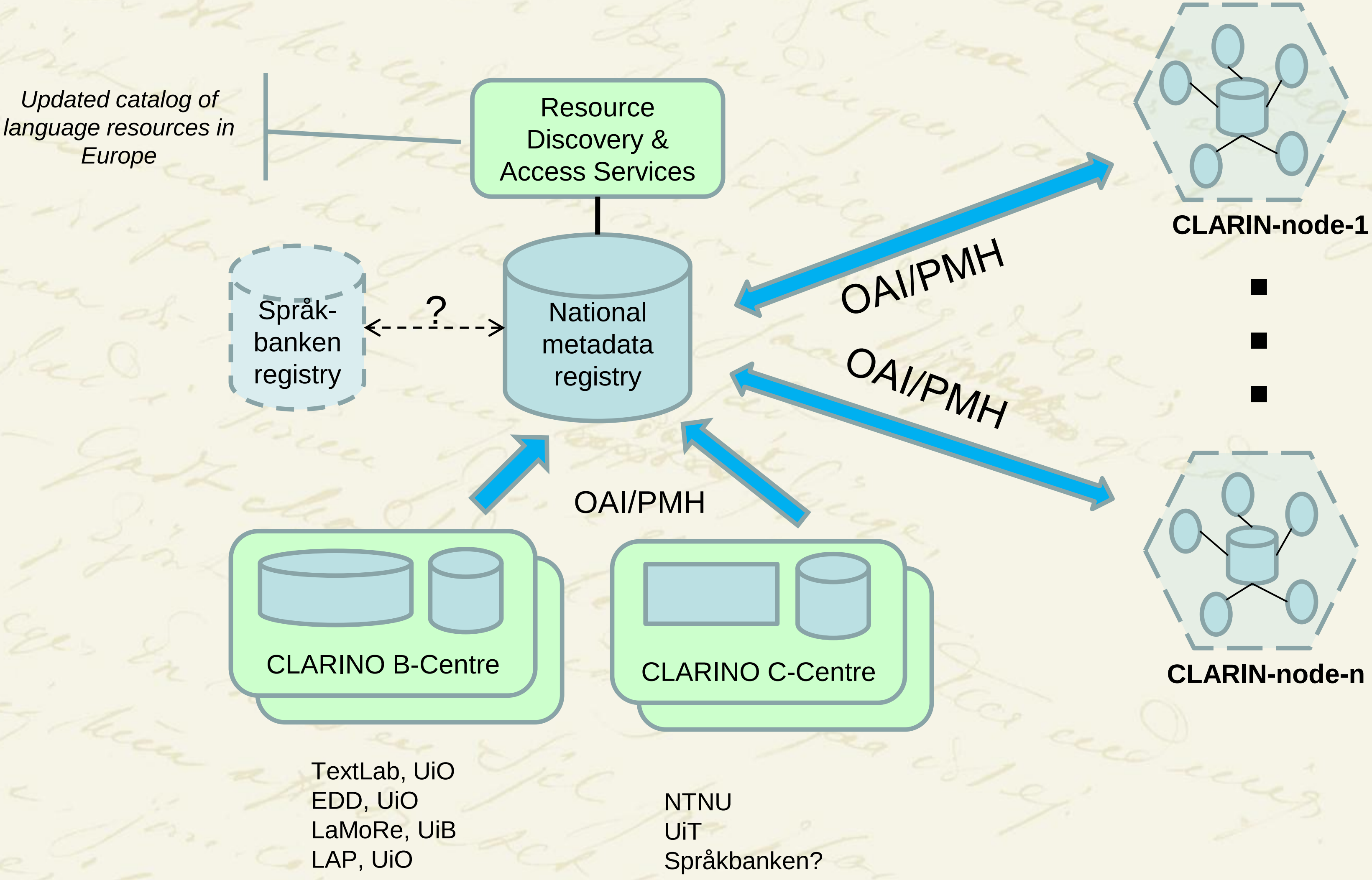
Language resources

. . .

Language resources

Nasjonalbiblioteket

Metadata exchange in CLARINO



Updated catalog of language resources in Europe

Resource Discovery & Access Services

Språk-banken registry

?

National metadata registry

OAI/PMH

OAI/PMH

OAI/PMH

CLARIN-node-1

■
■
■

CLARIN-node-n

CLARINO B-Centre

CLARINO C-Centre

TextLab, UiO
EDD, UiO
LaMoRe, UiB
LAP, UiO

NTNU
UiT
Språkbanken?

Nasjonalbiblioteket

Metadata infrastructure requirements (1) – the metadata model

Expressivity

- Cover needs of all target user groups and all resource types
- Must include preservation data (which?)
- Operate with the concept of *collection*
- Relationships between resources part of metadata
  - including part-of

Conceptually sound

- Flexible and extendible

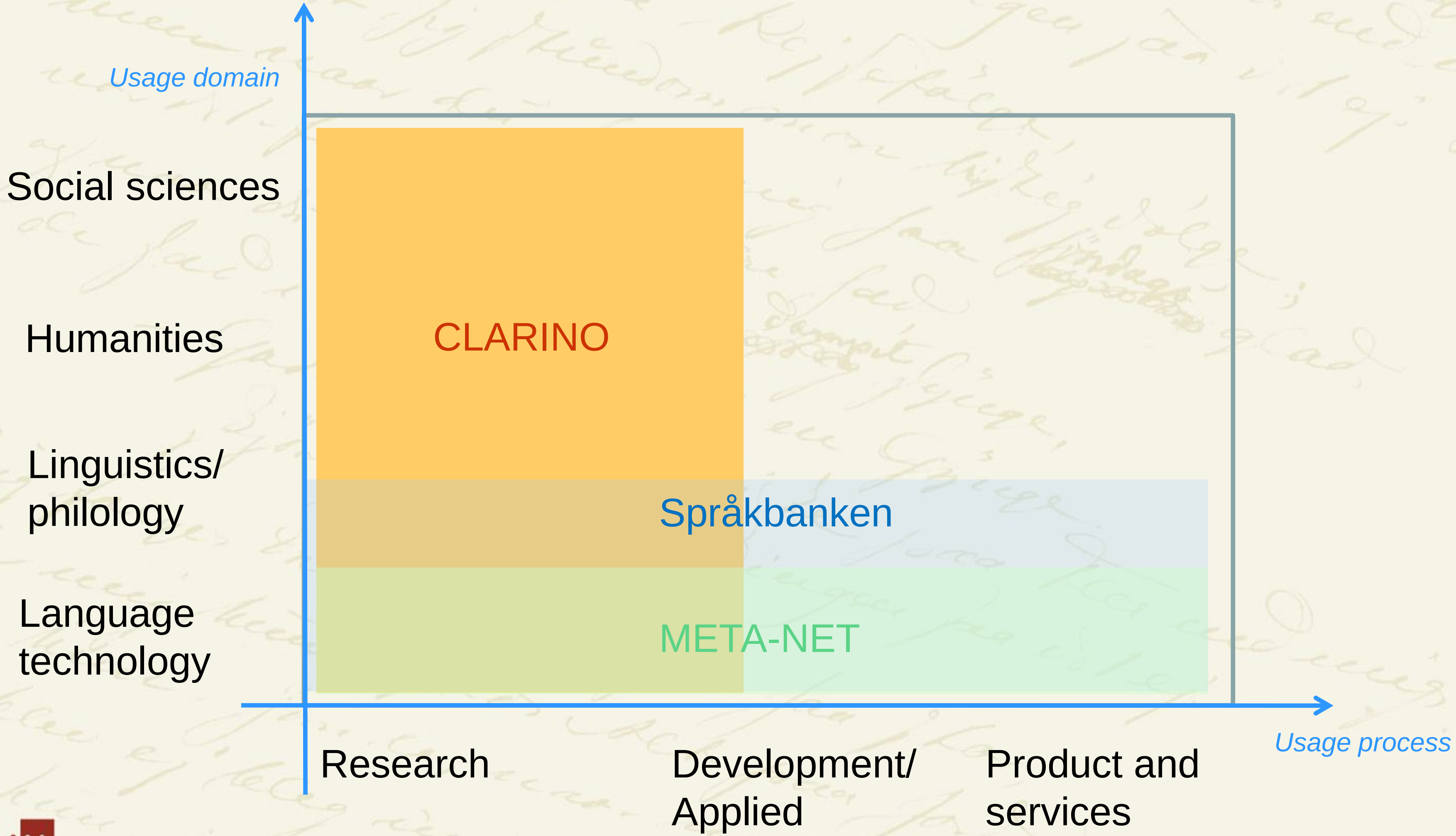
Community acceptance

- Copy-cat approach

Internal model vs. exchange model

- Mappability between chosen model and others like DC, OLAC, schemes used in legacy metadata

Target groups



Metadata infrastructure requirements (2) – end user facilities

Discover resources

- Search and browse facilities
  - Refinement of retrieved result sets
  - Various visualisations of result sets

Access/Acquire resources

- Getting the actual content

Personalisation?

(Pragmatism important) – basic things first!

Metadata infrastructure requirements (3) – management facilities

Metadata ingest methods

- Metadata editor (manual ingest)
- Import facilities – harvesting from other CLARINO centres and other CLARIN nodes

Metadata processing

- Analysis facilities
- Transformation facilities

Metadata dissemination

- Export facilities, e.g citation format, rdf/linked data,
- OAI/PMH gateway to other CLARIN nodes

Alternative solutions – as we see it

Existing infrastructures for language resources:

- META-SHARE
- CMDI (CLARIN)

(Using existing metadata systems for bibliographical resources)

(Develop from scratch)

CMDI – Clarin Metadata Infrastructure

General constituents:

- A registry (ISOcat) of data categories (i.e. metadata elements), to be grouped into components (recursively) which again might be grouped to form profiles
- A component registry (of components and profiles)
- A relation registry (early version)

Currently only a framework, not a deployable solution – as it is

Current set of metadata elements needs more curation

- Still somewhat experimental
- No agreement on standard basic set of metadata elements
- Needs metadata modeler to specify a suitable profile

Some prototype realisations (real infrastructures) exist

META-SHARE

Developed in an EU FP7 project

Target group : Language technology research & development

More or less complete infrastructue (when finished)

Metadata model fully specified – no metadata modeller needed

- Minimal and maximal element set
- Based on fixed resource typology
- From workshop in August 2011: Seems to correspond well to what Språkbanken's reserach users need.

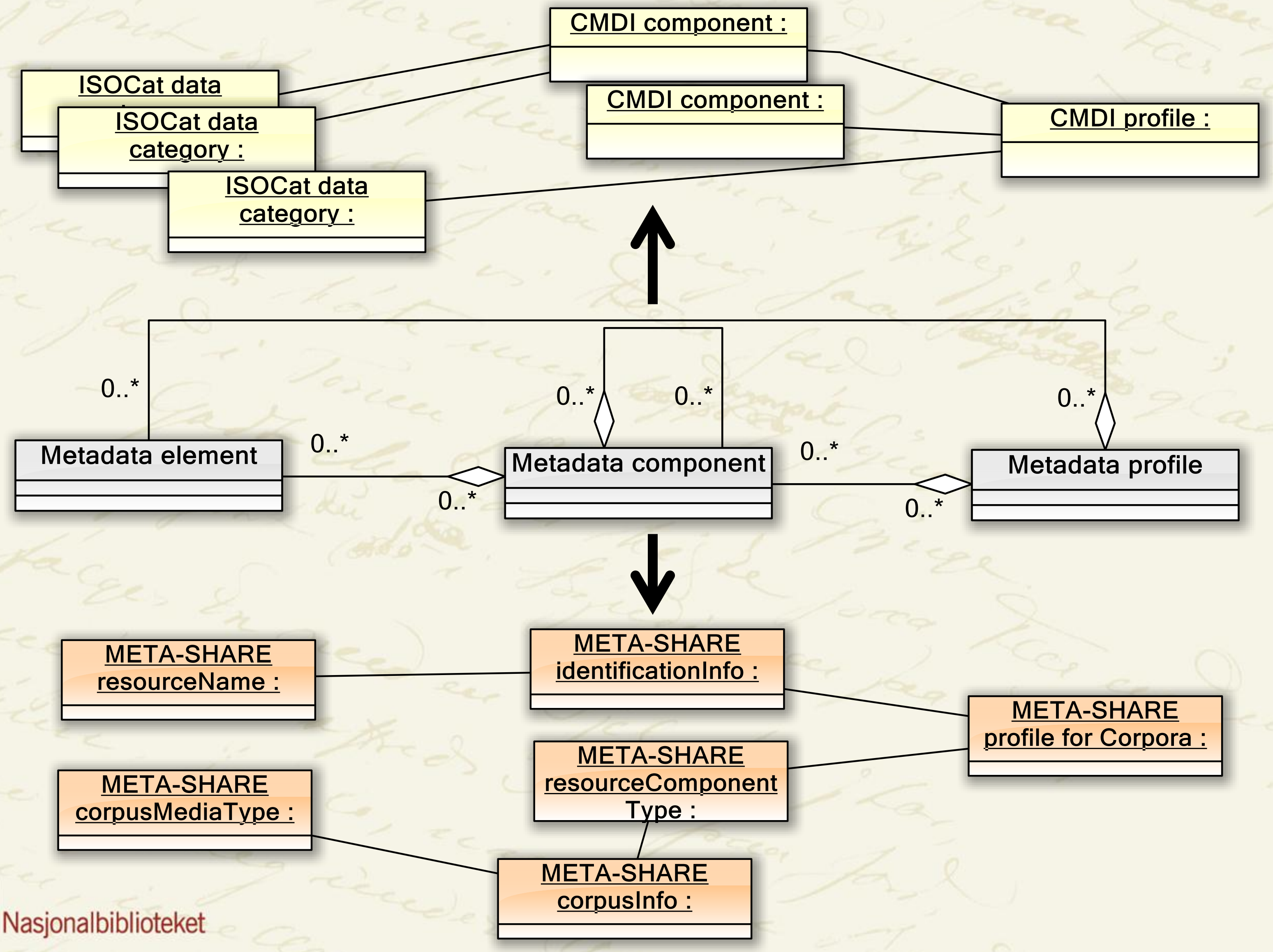
Work in progress – some important things are still lacking

- OAI/PMH support
- Mapping facilities
- Parts of user interface

Next version will be the first public version

Will be open source (downloadable from GitHub)

CLARIN – metadata component model



Nasjonalbiblioteket

Actions ahead

Install and populate META-SHARE

Provide for systematic user feedback

- on metadata model
- on system facilities

Consider how to remedy shortcomings

- Extend/modify META-SHARE software? (Open source)
- Induce META-SHARE to be explicitly included in the CMDI framework

Open issues – future ponderings

Metadata registry of Språkbanken vs. CLARINO

- Nature of interoperability must be clarified

How to connect LRs to bibliographical resources?

Where does search in metadata end and search in content/resources begin?

- Cf. metadata based search in Glossa and Korpuscle

How to obtain high quality metadata?

- Depends on many things, but not least on the description practice