Setting up a CLARIN centre

Dieter Van Uytvanck

Newcomers' Workshop

4 December 2018

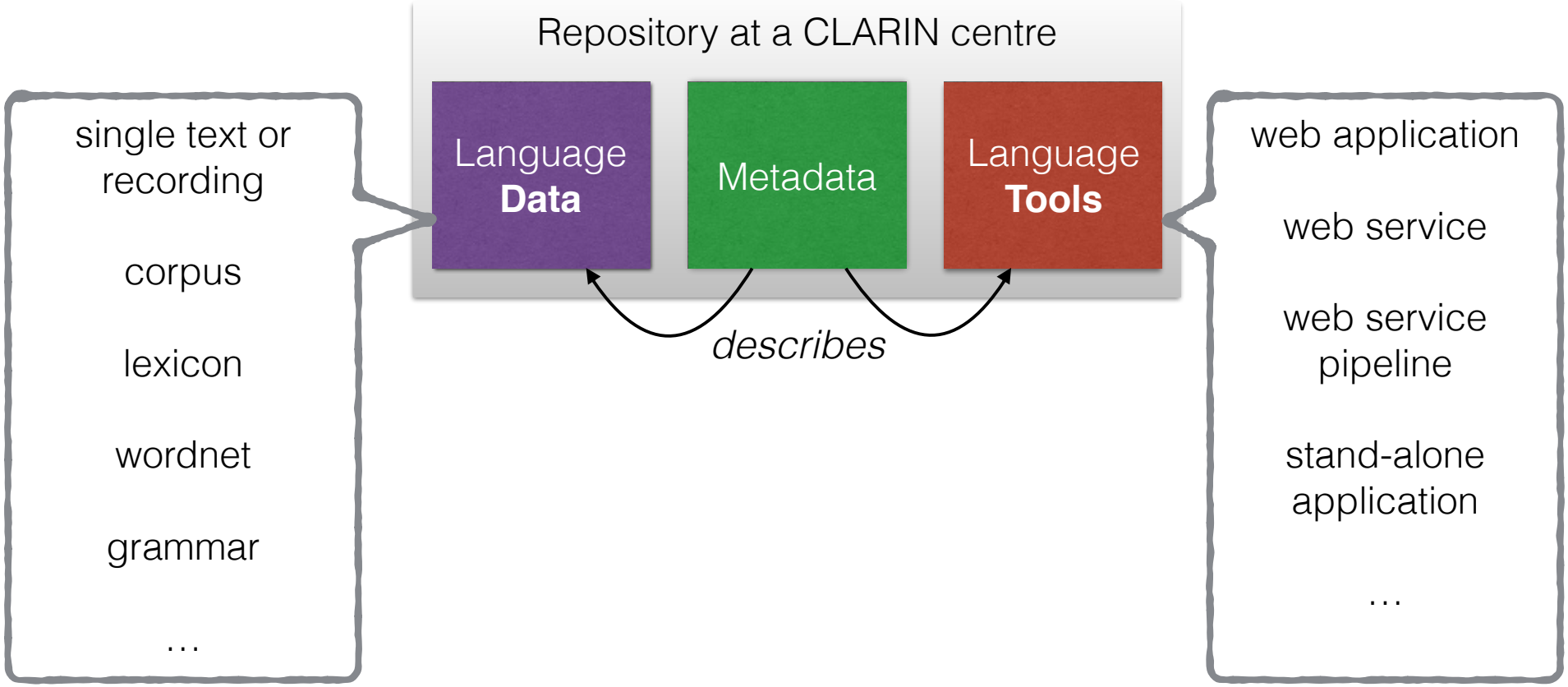
Utrecht

Overview

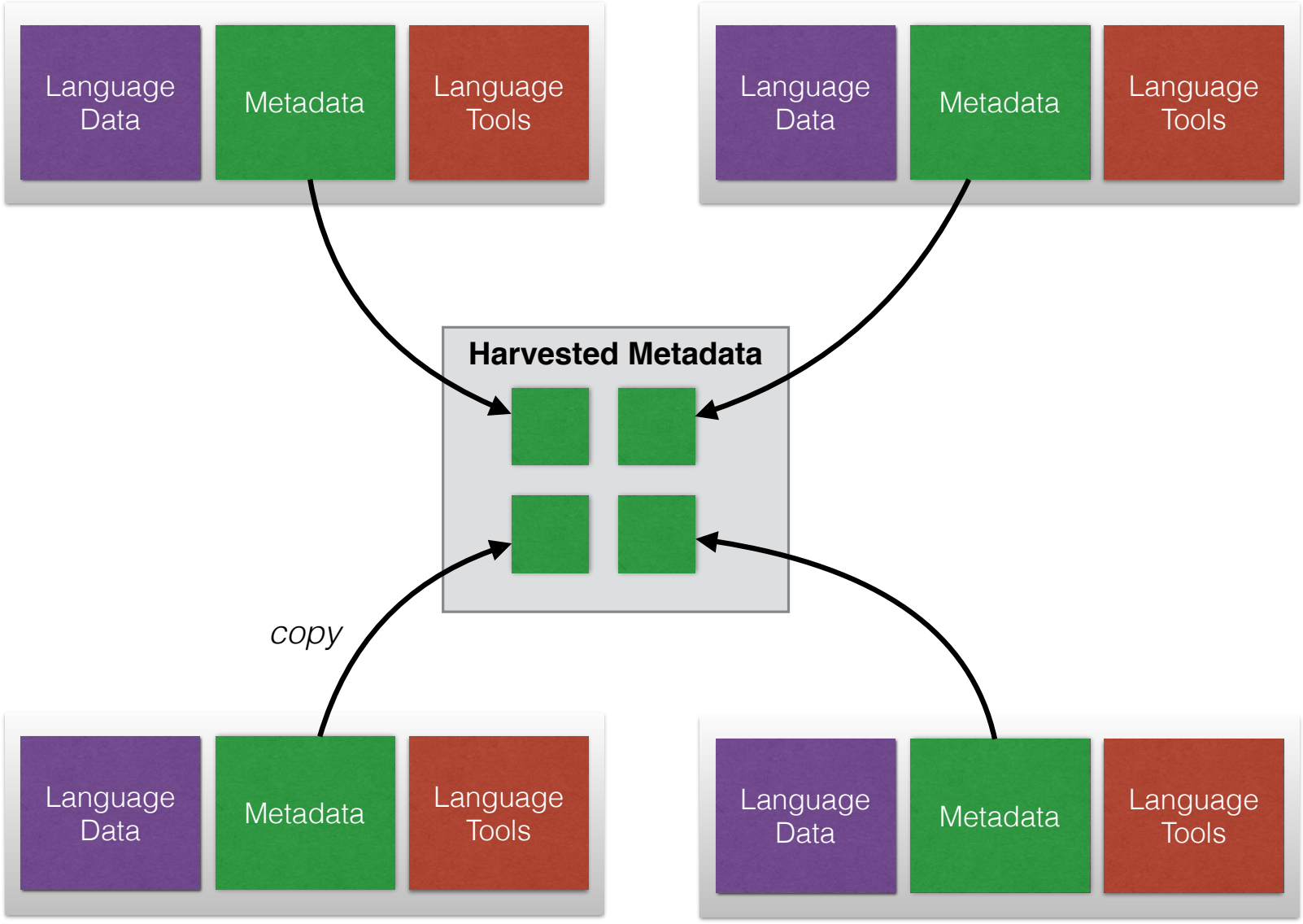
- The CLARIN technical infrastructure from a bird's eye
- Infrastructure pillars:
  - Repositories
  - Metadata & VLO
  - Tools & LR Switchboard
  - Federated Login
  - Federated Content Search
- CLARIN centres:
  - Types of centres
  - B-centre assessment
    - Requirements
    - Core Trust Seal
    - Procedure

The CLARIN technical infrastructure from a bird's eye view

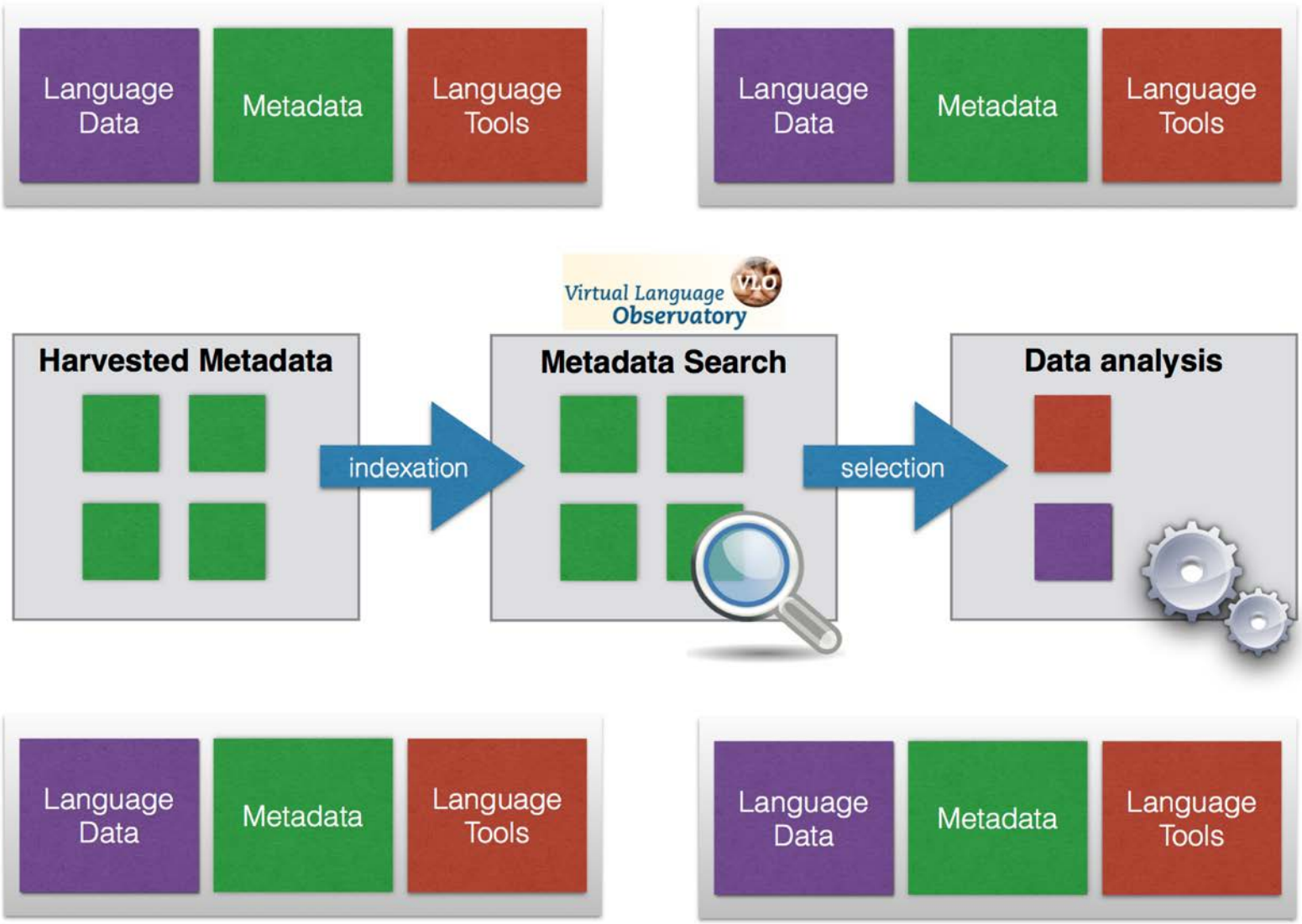
Architecture: Repositories



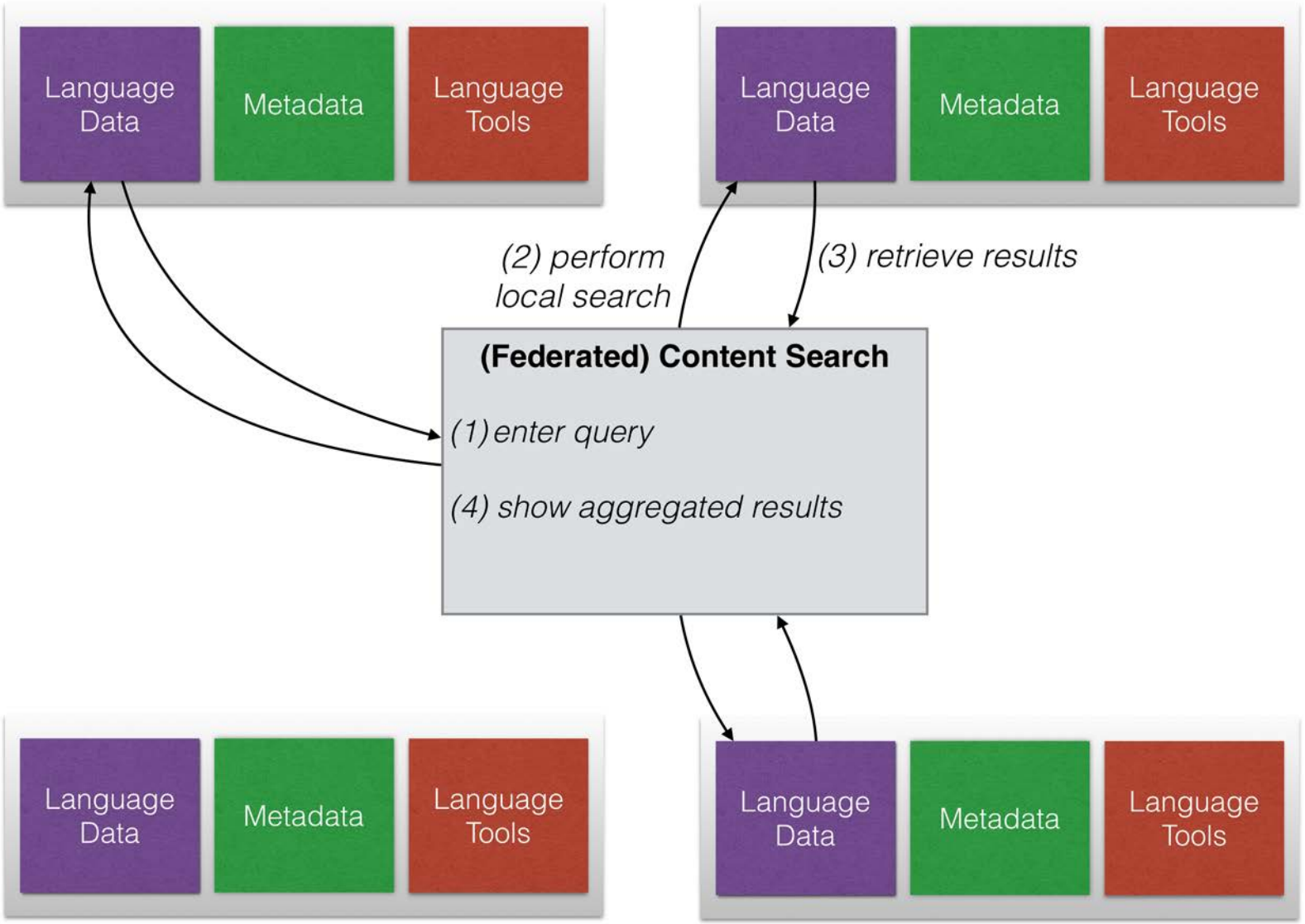
Architecture: Harvesting



Architecture: Processing



Architecture: Federated Content Search

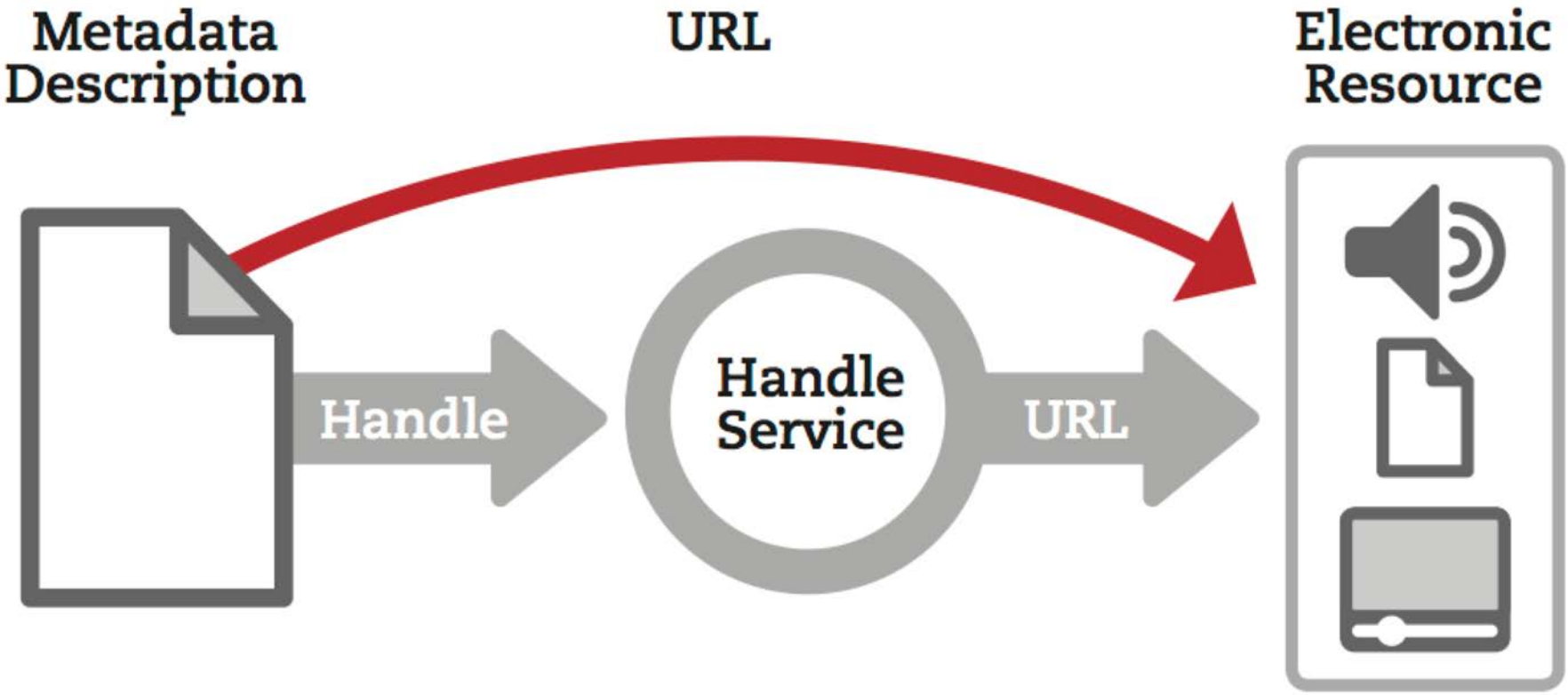


Summing up

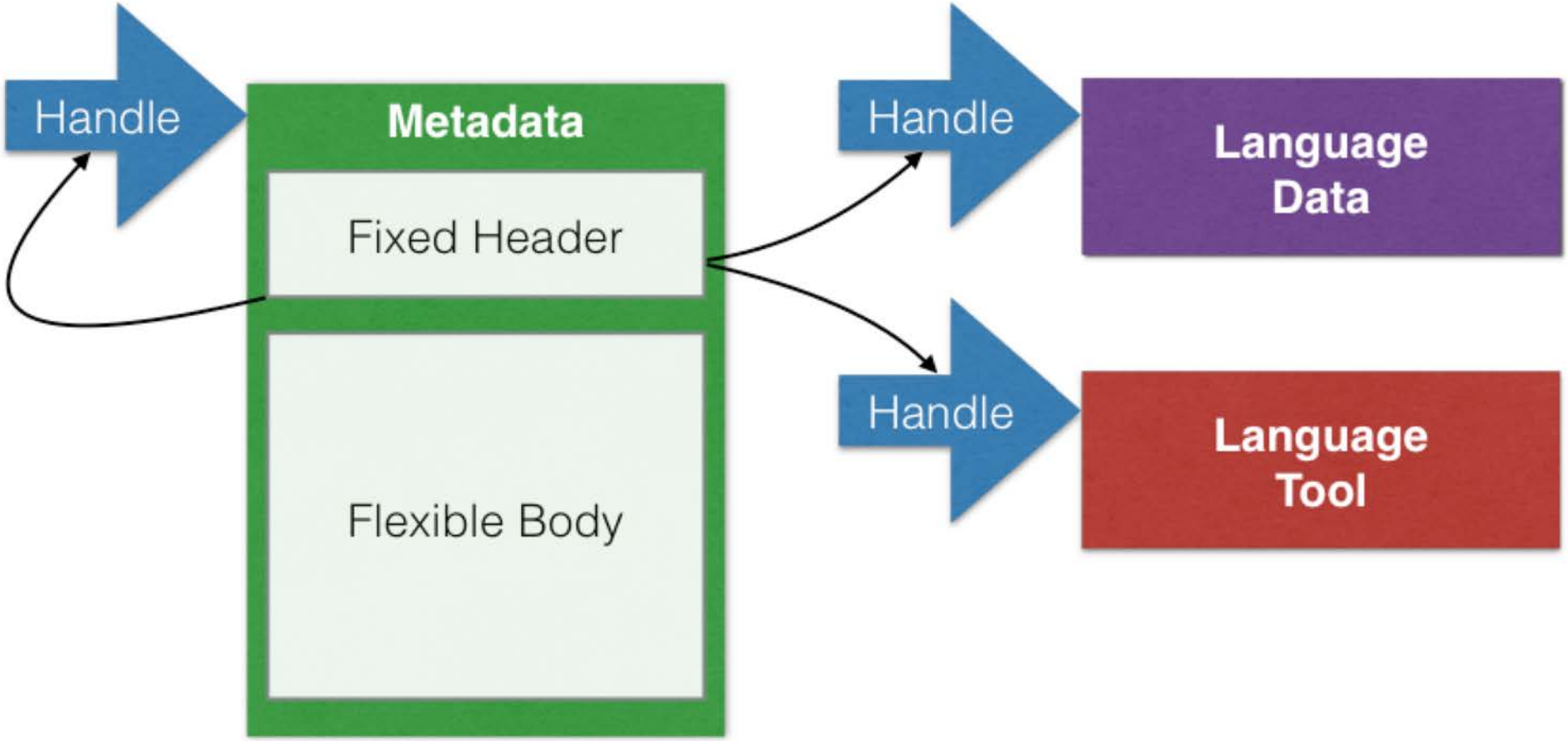
- Basic principles:
  - compatibility on protocol and format level
  - acknowledging the unique situation of each centre (history, organization, technology, etc.)
  - focus on the strengths that a centre can contribute
- No obligation to use specific software stacks
  - Of course it can save resources to reuse existing solutions
- Striking the right balance between do-it-yourself and reuse is one of the most important steps in the process of becoming a CLARIN centre

Infrastructure pillars

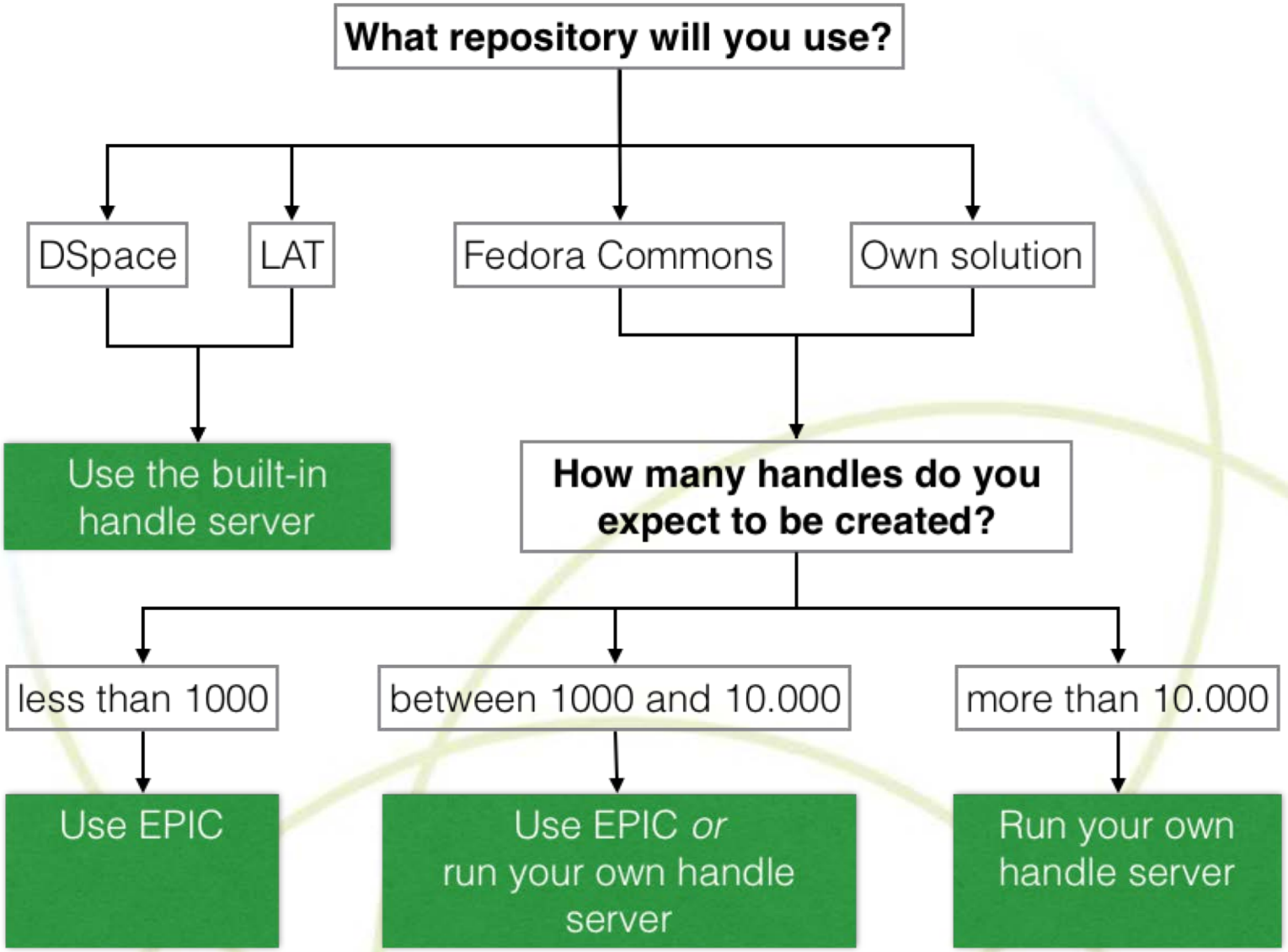
Persistent Identifiers (PIDs)



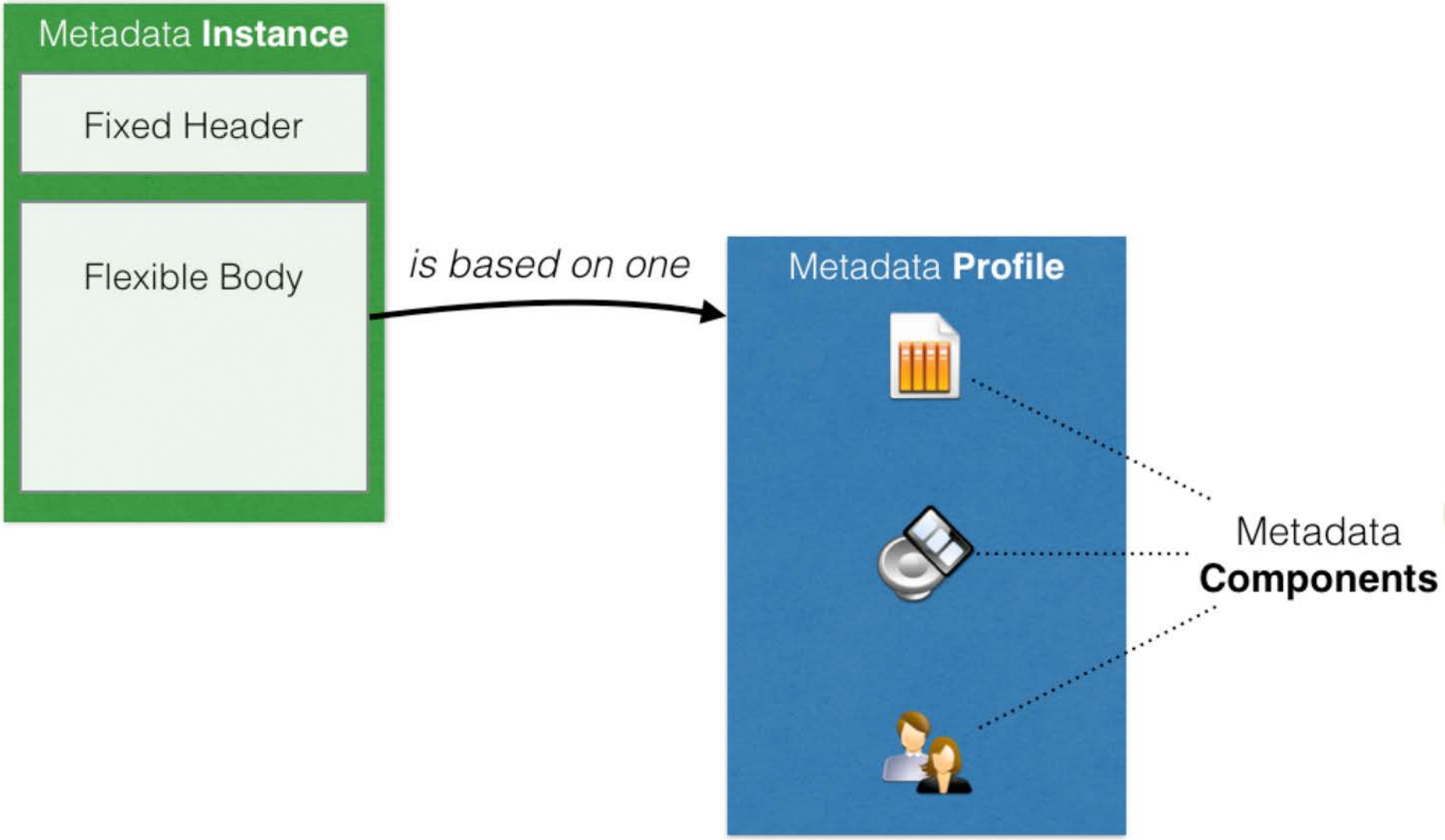
Persistent Identifiers (PIDs)

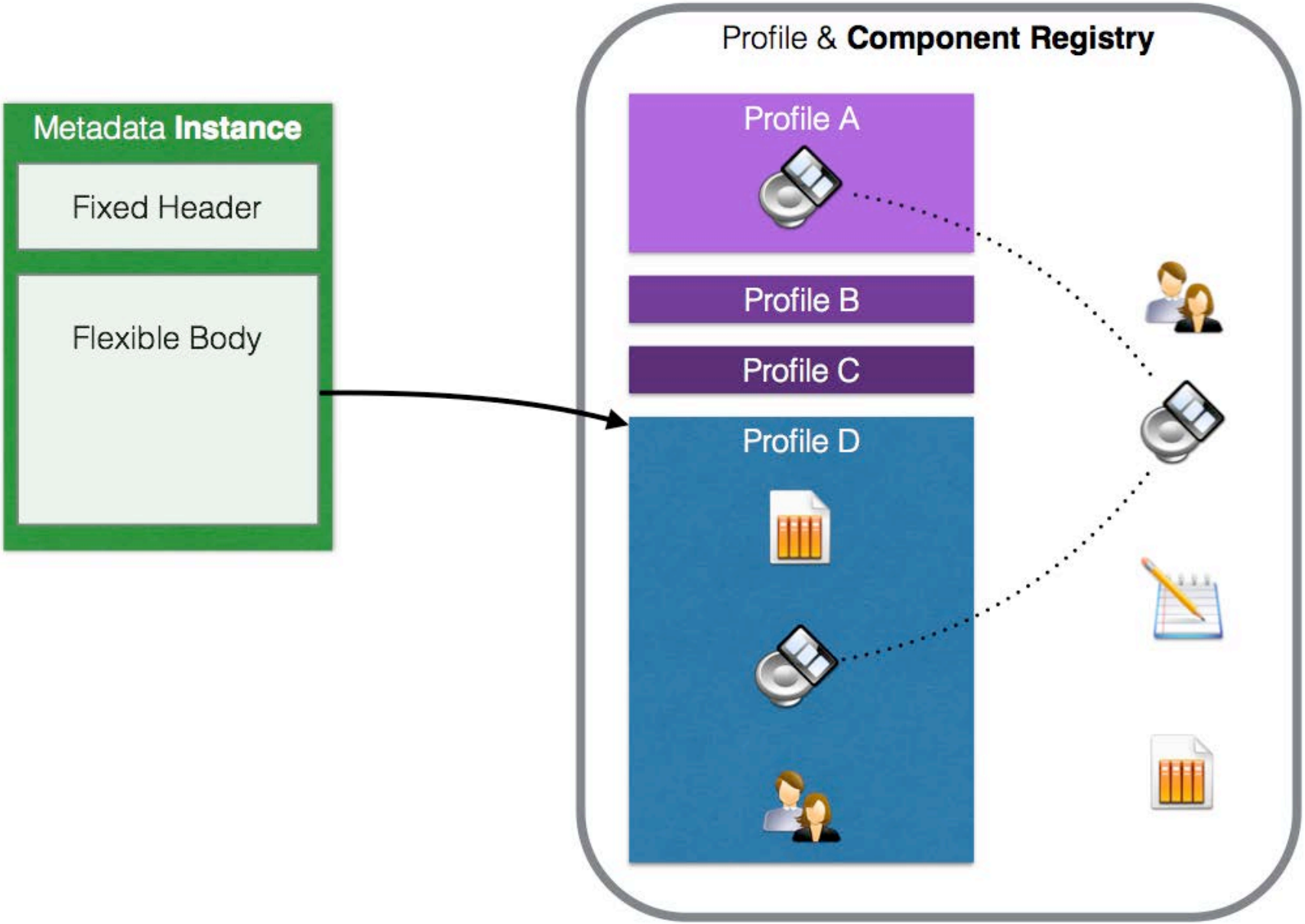


Persistent Identifiers (PIDs)



Component Metadata (CMDI)




Metadata **Instance**
Fixed Header
Flexible Body
Profile & **Component Registry**
Profile A
Profile B
Profile C
Profile D

Interview Profile

General information
Component

- **Title**: text
- **Creation Date**: date

Sound recording
Component

- **Format**: wave | mp3
- **Length**: number

Actor
Component

- **First name**: text
- **Last name**: text
- **Birth Date**: date
- **Role**: interviewer | interviewee

Concept Registry

definition of:

- Title
- Creation Date
- Format
- Length
- First name
- Last name
- Birth date
- Role
- wave
- mp3
- interviewer
- interviewee
- …

CMDI: Basic ideas behind it

- Allowing for the flexibility needed: many different subcommunities have their own wishes to provide detailed metadata descriptions
- Stimulating re-use: most providers should be able to re-use an existing profile
- Provide a standard way to
  - Refer to
    - digital objects (in the fixed header)
    - landing pages
    - search pages
    - search services
  - Express hierarchies in metadata files

CMDI: What we currently have in place

- 177 profiles, 1255 components at https://clarin.eu/componentregistry
- Over 1900 concepts registered at https://clarin.eu/ccr
- Over 22 CLARIN centres providing native CMDI metadata
- Important conversion workflows in place:
  - Europeana Data Model (775,000 records)
  - OLAC & Dublin Core
  - MODS
  - TEI headers
  - under investigation & in preparation:
    - DDI (social sciences)
    - DataCite metadata
- Catalogue of harvested metadata: https://vlo.clarin.eu

CMDI: Must reads & myths

- CMDI best practice guide: https://www.clarin.eu/content/cmdi-best-practices-guide
- Metadata in CLARIN – the FAQ: https://www.clarin.eu/faq-page/267
- Myth: CMDI must be used as backend format in my repository.
  - Incorrect! The only requirement is to deliver it via OAI-PMH.
- Myth: CMDI records must be created manually with an editor
  - Incorrect!
    - A simple and automatic conversion from your existing metadata format or database can be sufficient.
    - For large amounts of metadata, manual editors are inconvenient.

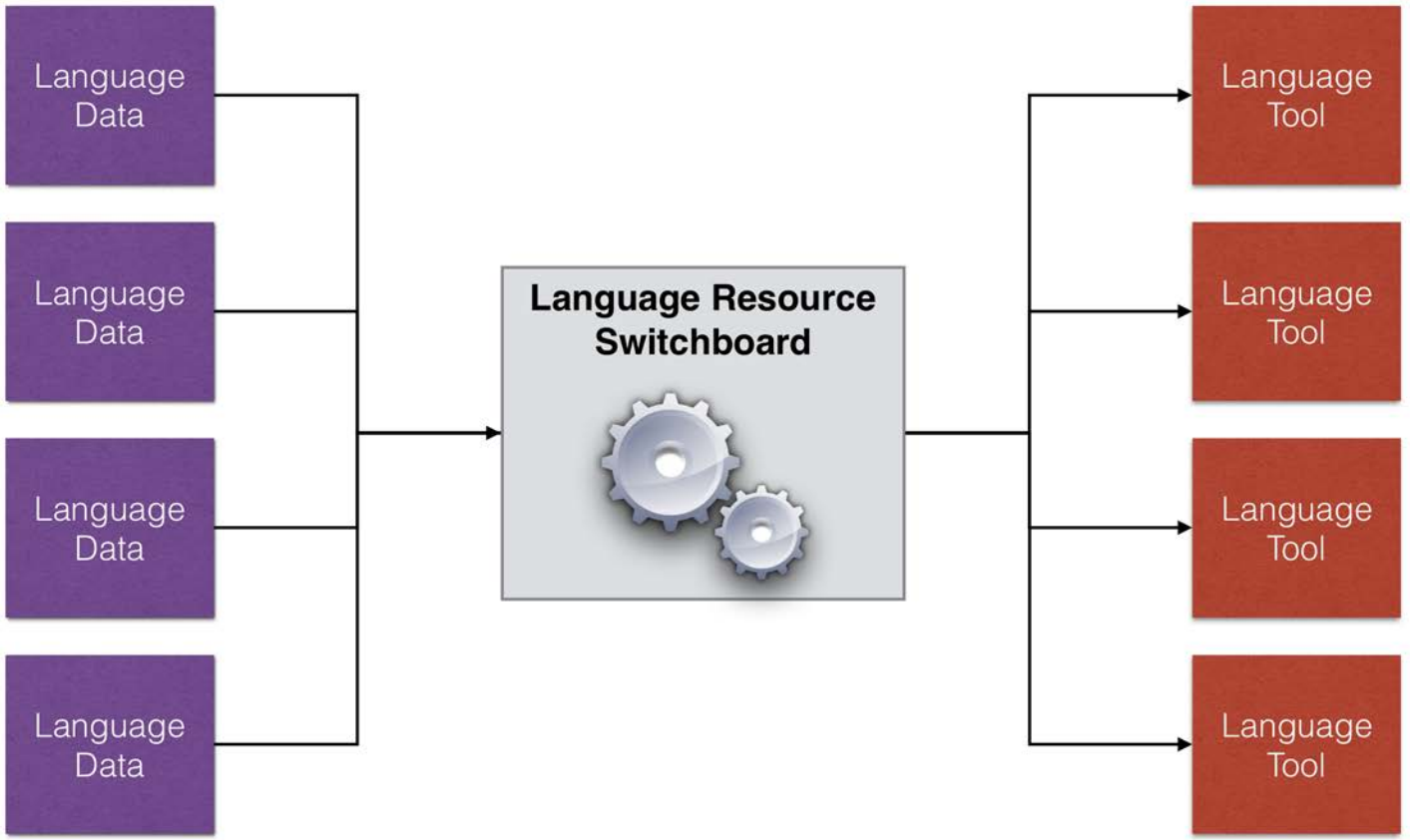
Repositories

- https://www.clarin.eu/content/repositories
- Off-the-shelf:
  - LINDAT-Dspace
  - Dataverse (for a C-centre, not fully B-centre ready)
    - But it is open source…
- Half-products
  - META-share repository + CLARIN-specific adjustments
  - Islandora (Fedora Commons + Drupal)
    - As used at HZSK and FLAT (TLA, Meertens-HuC)
- From scratch
  - Based on Fedora Commons
  - Based on your existing home-built repository

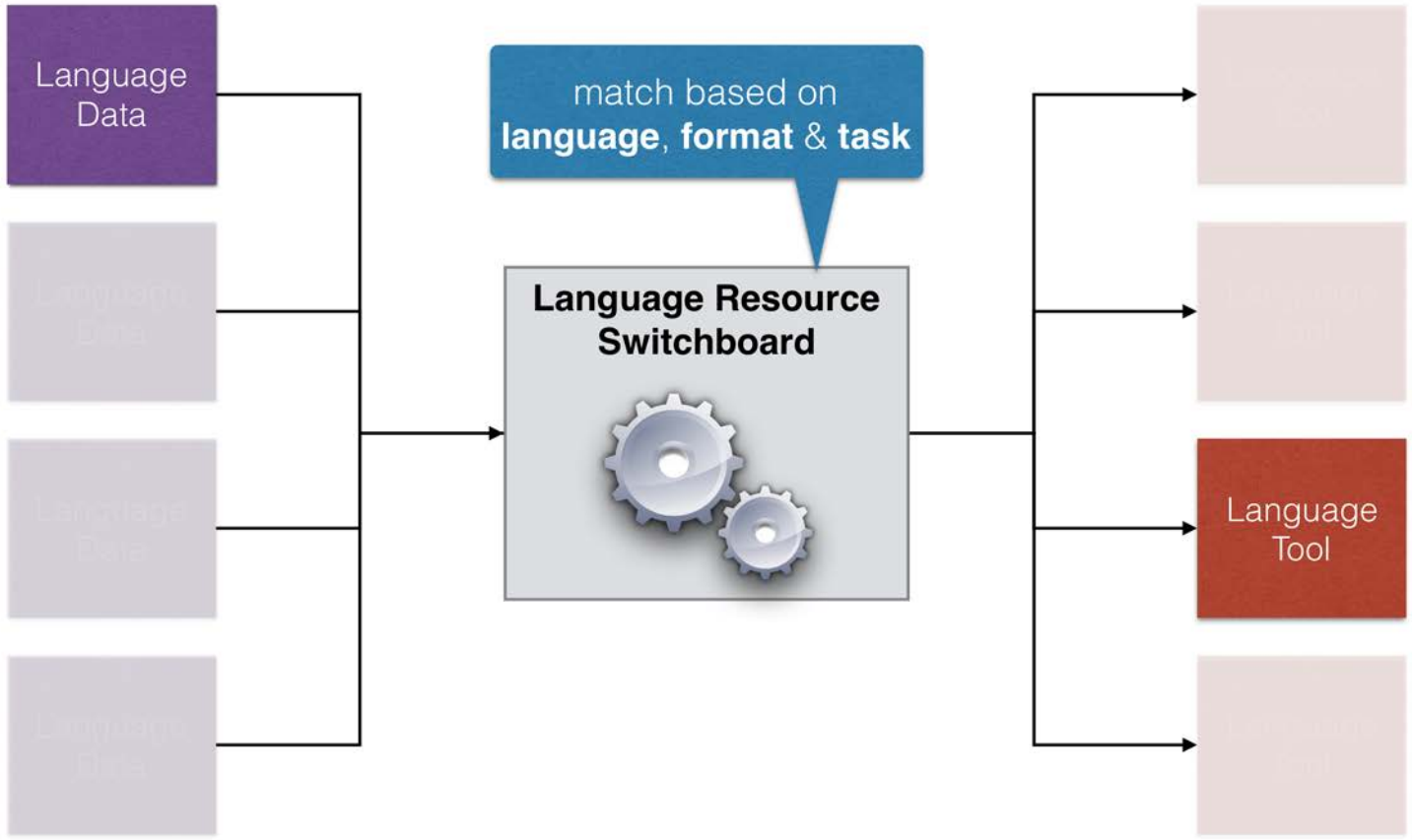
Repositories: learn more

- https://www.clarin.eu/event/2017/clarin-plus-workshop-facilitating-creation-national-consortia-repositories
- https://www.clarin.eu/event/2016/clarin-workshop-dspace-digital-repository
- Lindat-DSpace tutorial:
  - https://www.youtube.com/playlist?list=PLlKmS5dTMgw3lJ4TffnBJOhprLd_-20jd

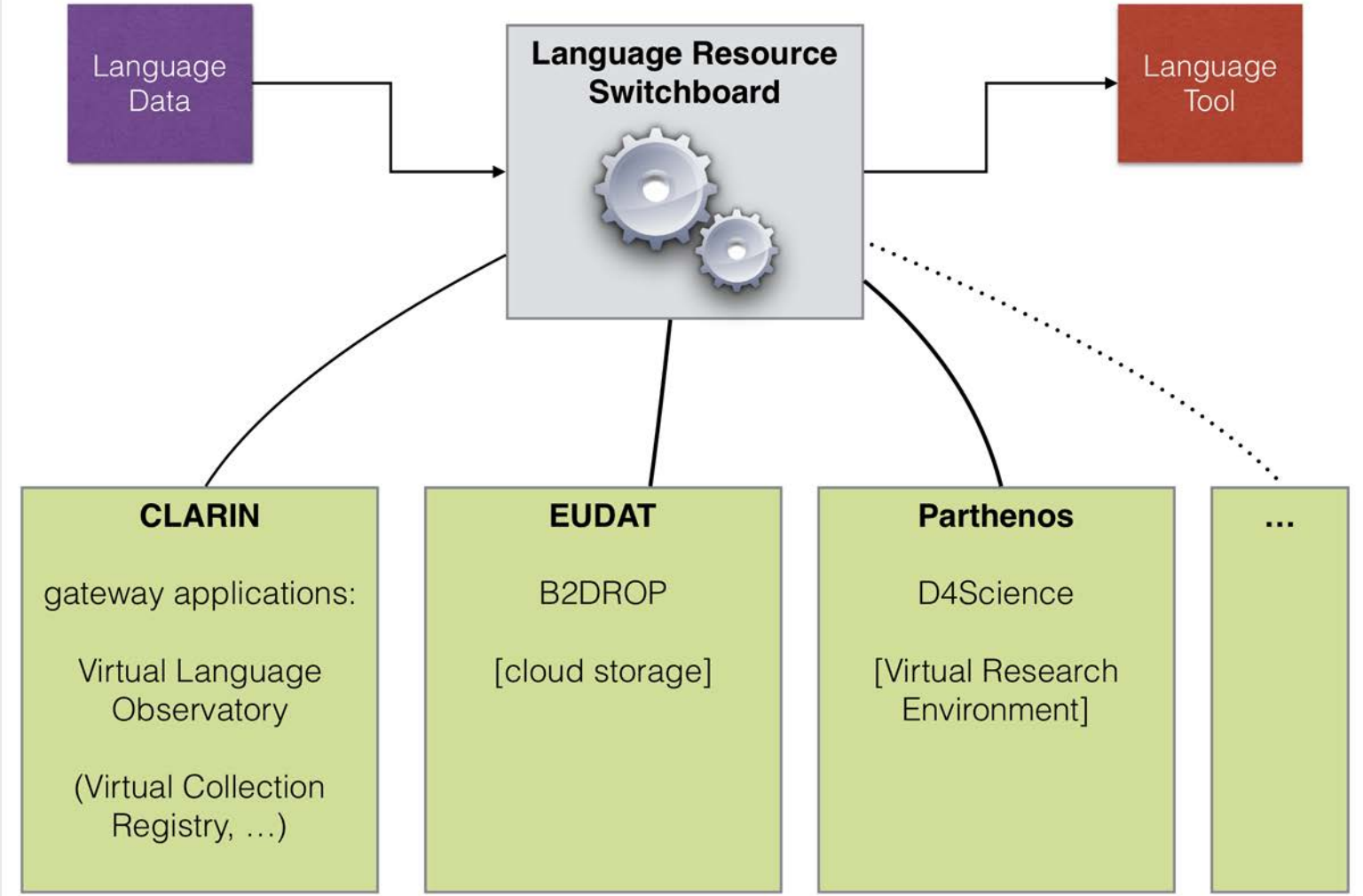
Tools & LR Switchboard



Tools & LR Switchboard



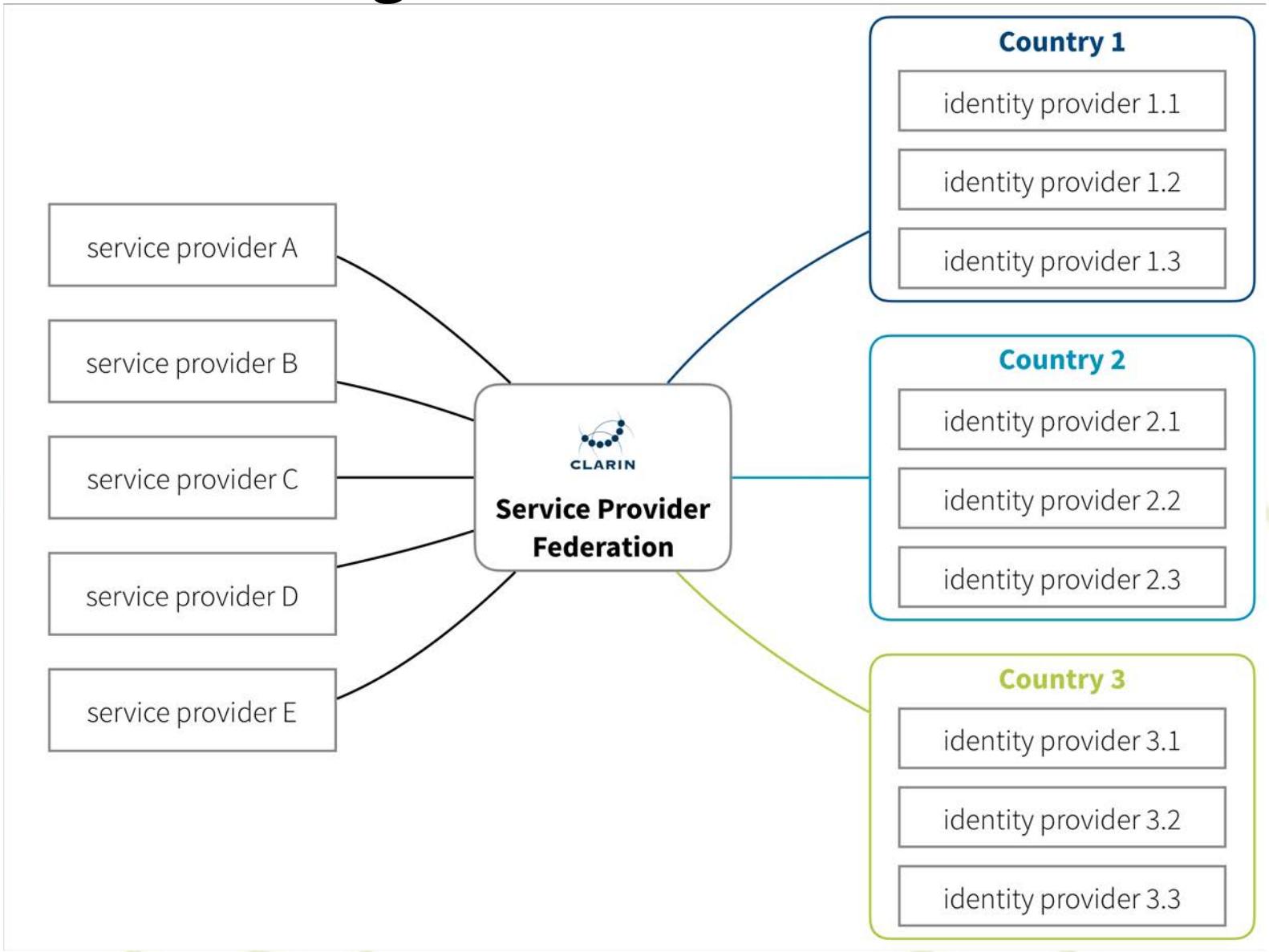
Tools & LR Switchboard



Tools & LR Switchboard: learn more

- Connecting a web application:
  - https://switchboard.clarin.eu/ > For Developers
- Use cases
  - https://office.clarin.eu/v/CE-2018-1196-language_resource_switchboard_use_cases.pdf
- Demonstration case:
  - https://www.clarin.eu/showcase/eosc-portal-demonstration

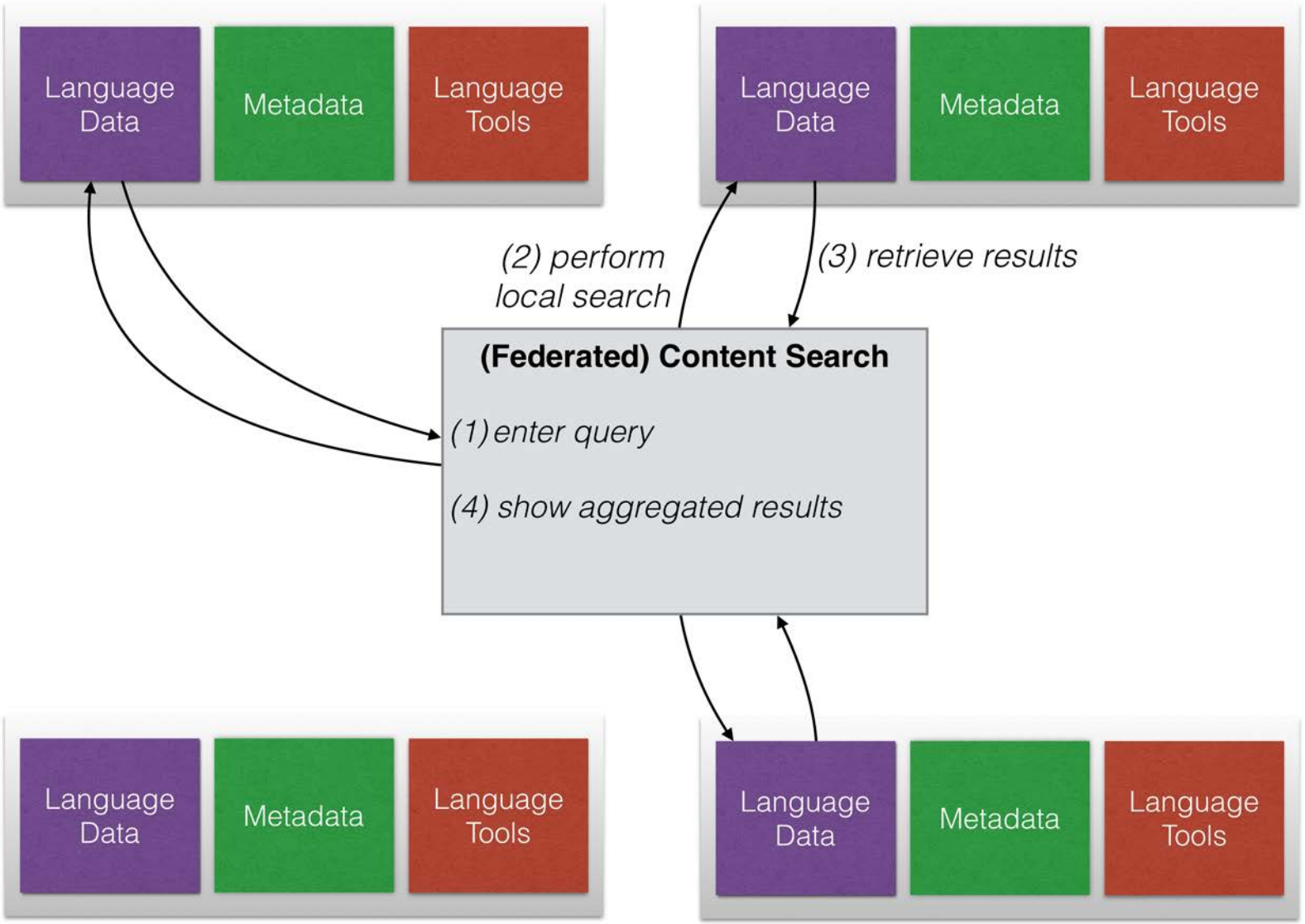
Federated Login



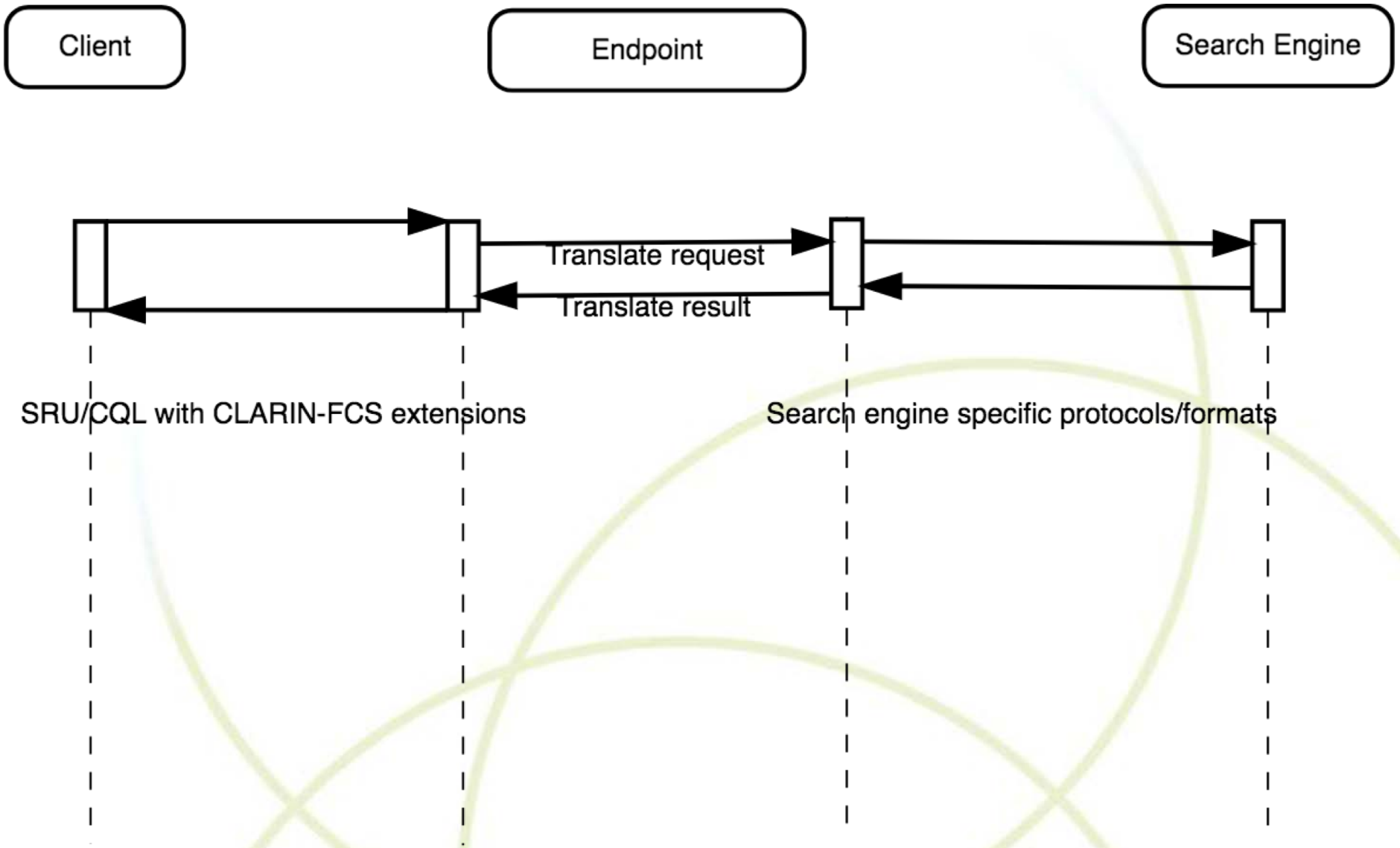
Federated Login

- Reading material:
  - Background (good starting point)
    - History and functioning of the Service Provider Federation: https://office.clarin.eu/v/CE-2017-1014-CLARINPLUS-D2_7.pdf
  - Overview (e.g. to get the paperwork running):
    - https://www.clarin.eu/spf
  - Technical instructions on creating and integration of a Service Provider:
    - https://www.clarin.eu/content/creating-and-testing-shibboleth-sp

Federated Content Search



Federated Content Search



Federated Content Search

- Specifications and background information:
  - https://www.clarin.eu/content/federated-content-search-clarin-fcs
  - https://trac.clarin.eu/wiki/FCS
- Some first notes on FCS for treebank searches:
  - https://www.clarin.eu/blog/blog-post-jan-niestadt-mini-workshop-korp-strix-and-blacklab-gothenburg
- Endpoint libraries on the way:
  - Korp
  - Kontext
  - NoSketchEngine
- More documentation to follow soon:
  - Technical documentation: December 2018 – January 2019
  - Outreach: January – February 2019

CLARIN centres

Centres: an introduction

- B-centres (***Service Providing Centres*** *aka* ***Certified Centres***)
- C-centres (***Metadata Providing Centres***, their metadata are integrated with CLARIN but they need not to offer any further services)
- K-centres (***Knowledge Centres***, part of the CLARIN Knowledge Sharing Infrastructure)
- E-centres (***External Centres*** offering central services without being part of any national consortium)

B-centres

- Assessment procedure description:
  - https://www.clarin.eu/node/3767

Centre registry

- Accessing information
  - https://centres.clarin.eu/
- Adding information:
  - https://www.clarin.eu/content/clarin-centres > Register a new centre
  - Adding an entry requires a CLARIN account
    - https://user.clarin.eu

Checklist

Misc links, developer resources

- clarin.eu/dev
  - Trac, especially Infrastructure Overview
- Slack: request access via trac@clarin.eu
- Mailing lists: all-centers
- Newsflash
- Piwik
- https://www.clarin.eu/applications
- https://curate.acdh.oeaw.ac.at

Thank you for your attention!

More information:

- www.clarin.eu

Feel free to contact

- our support addresses at https://www.clarin.eu/content/support
- me via dieter@clarin.eu