CLARIN
Common Language Resources and Technology Infrastructure

It's fun to play with the

Component **M**eta**D**ata **I**nfrastructure

CMDI: metadata for language resources à la carte

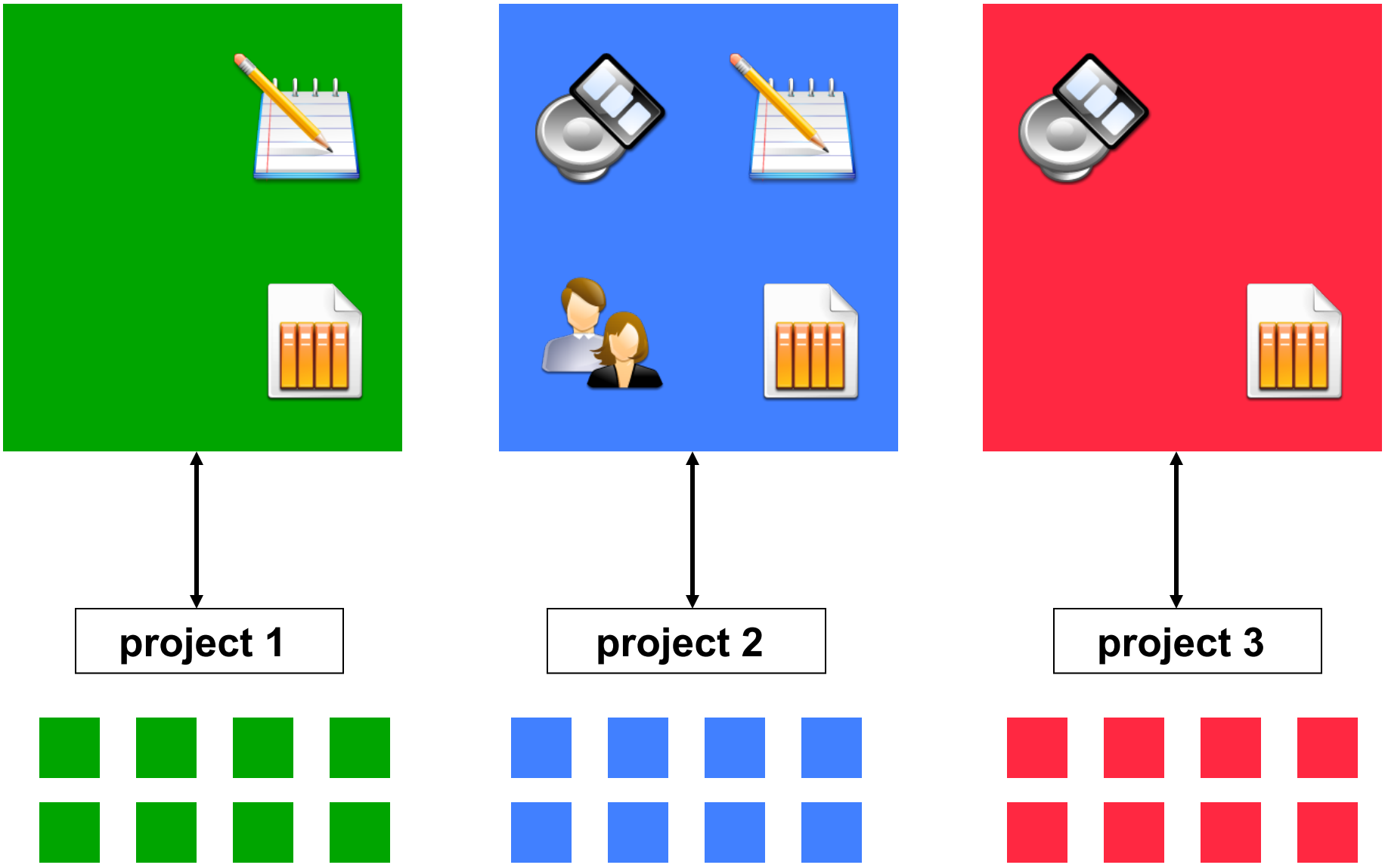
Dieter Van Uytvanck
Max Planck Institute for Psycholinguistics
Dieter.VanUytvanck@mpi.nl

MetaNord Metadata workshop
Oslo
2012-06-05

Overview

- Traditional metadata
- Component metadata
- Data categories
- The big picture
- In practice:
  - Building components
  - Using components

Traditional Metadata



Traditional Metadata: problems

- Lack of flexibility
  - Too many fields...
  - ... but not the ones I am looking for!
- Lack of interoperability
  - My metadata does not work with your infrastructure!
  - Nederland? Netherlands? The Netherlands? Holland? NL?

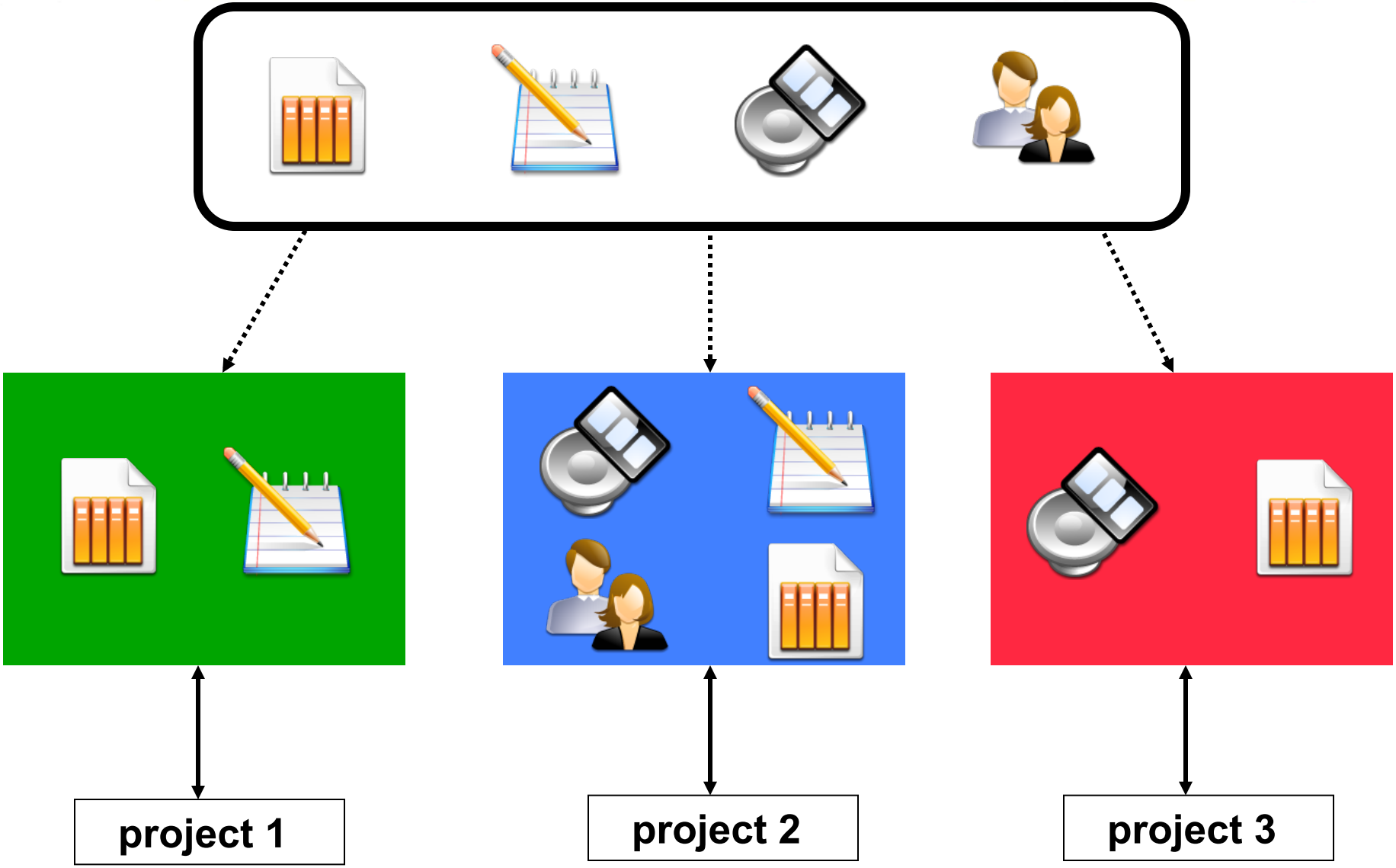
Context

- Other Metadata Infrastructures in our domain:
  - IMDI, OLAC/DC, TEI header, MetaShare
- Problems:
  - Inflexible: too many (IMDI) or too few (OLAC) fields
  - Limited interoperability
  - Problematic (unfamiliar) terminology for some sub-communities.
  - etc.

CLARIN Project - CMDI

- Metadata infrastructure based on a "Component Metadata Model"
- Aims
  - Flexibility
    - Researcher should themselves decide what metadata fits their needs
    - Offer ready made metadata components
    - Allow creation of new metadata components needed
  - Interoperability built-in
  - Complete Infrastructure: software for editing, harvesting, exploitation
  - Compatibility with existing frameworks: OLAC, IMDI

Component Metadata

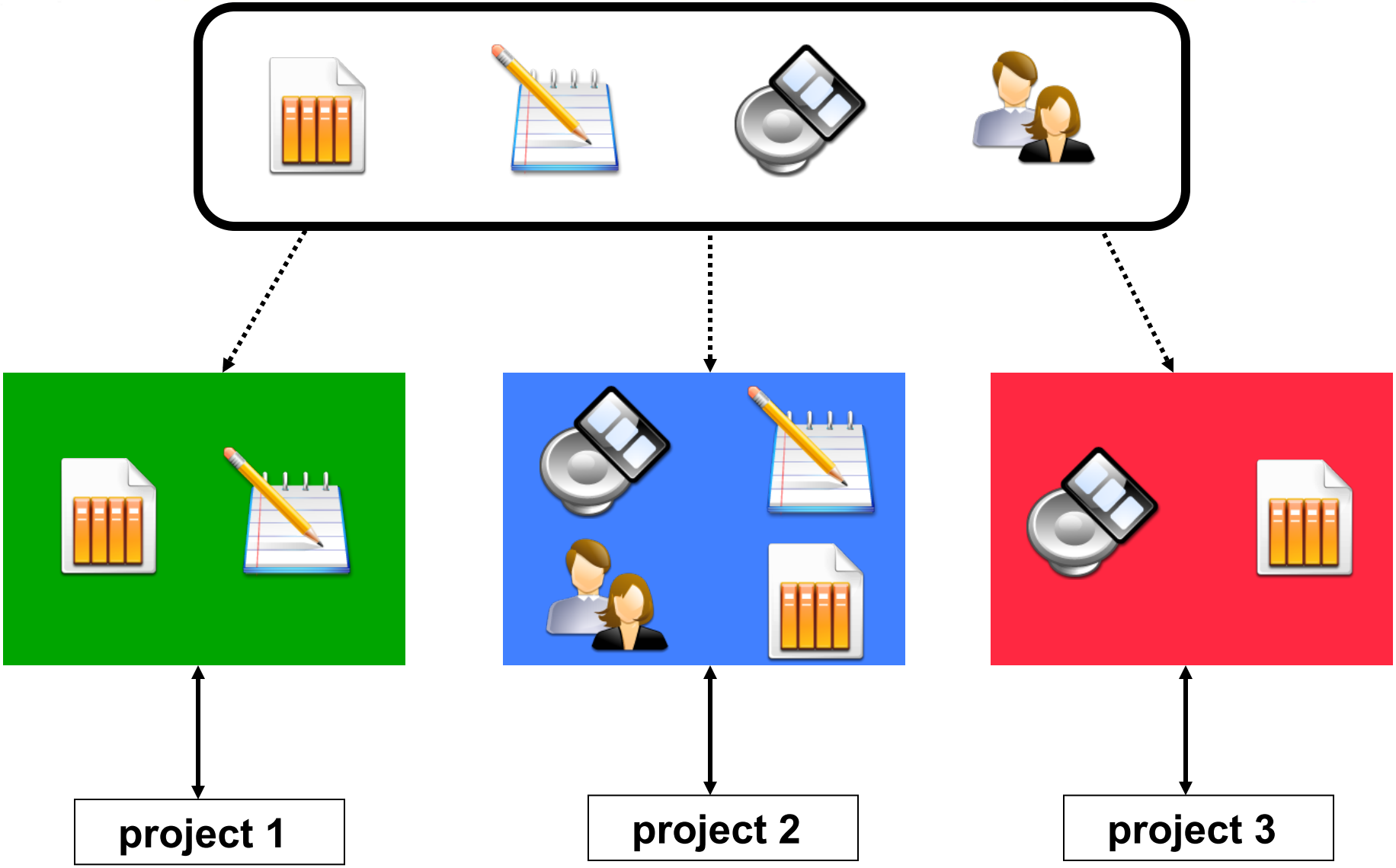


Some terminology

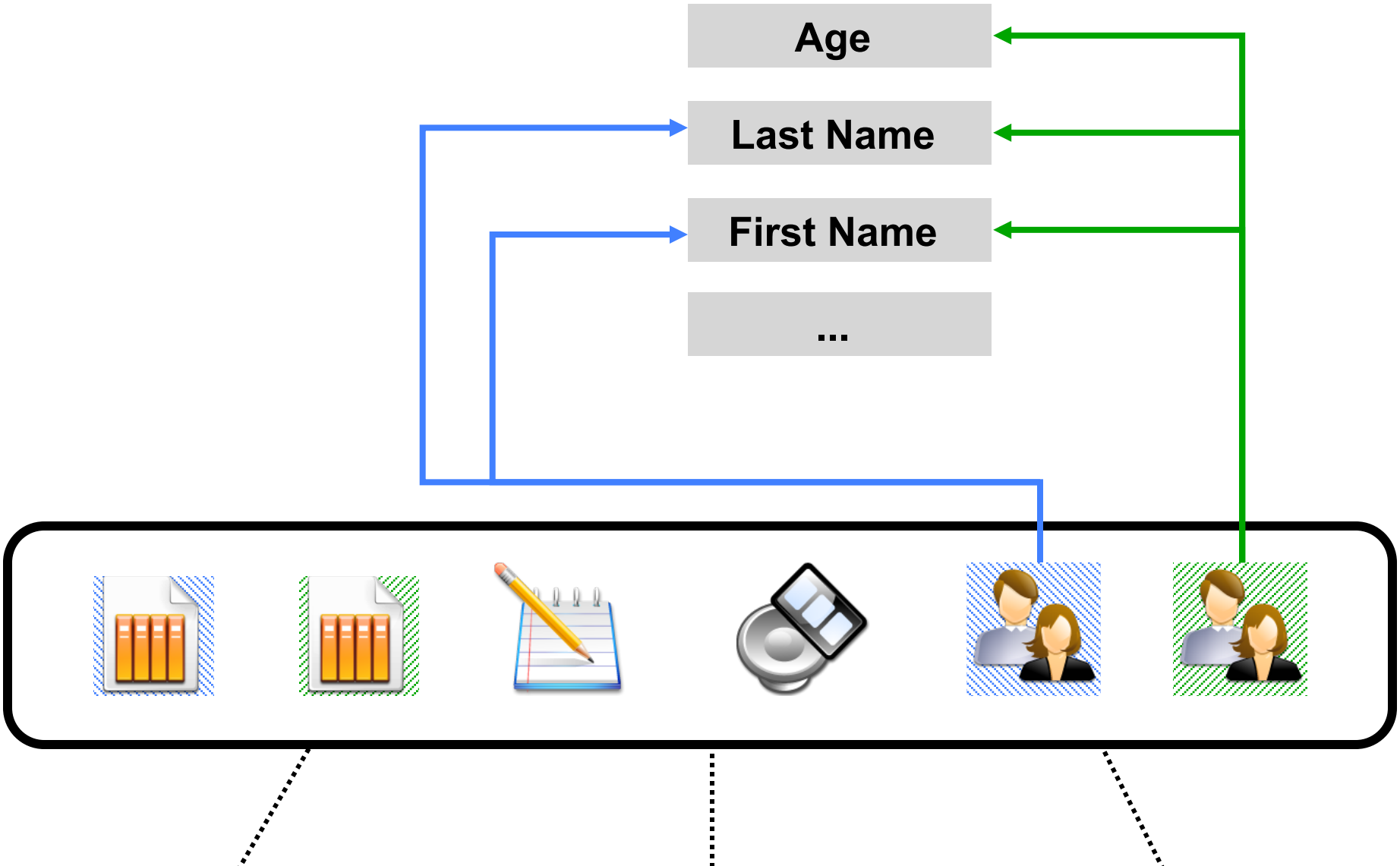
- **Element** = atomic unit (a “field”) – e.g. recording date
- **Component** = set of elements – e.g. Actor
- **Profile** = set of components – e.g. OLAC profile
- **Instance** = one metadata description – e.g. myresource.cmdi

- Schema = technical (formal) grammar describing a profile – e.g. olac.xsd

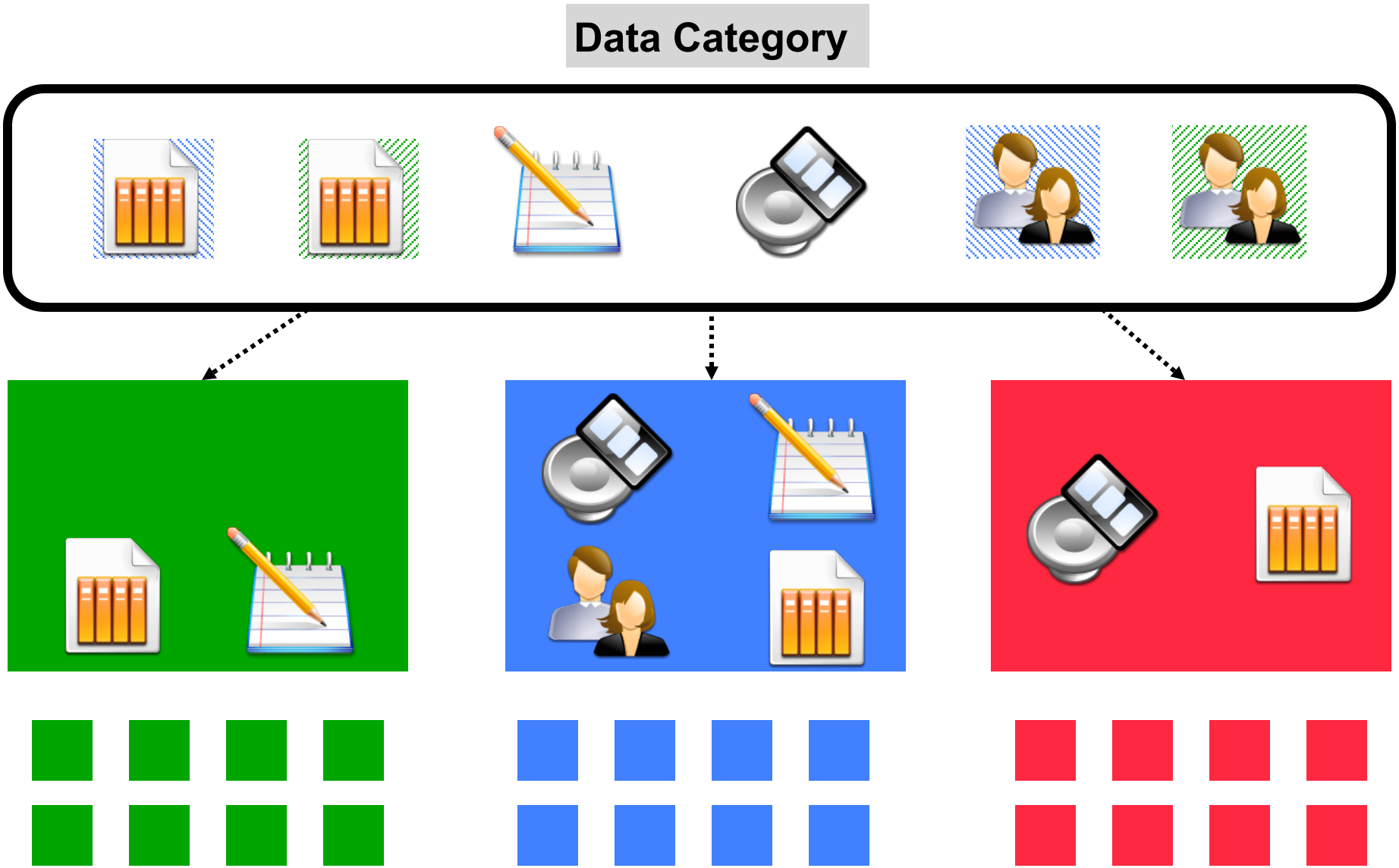
Data Categories



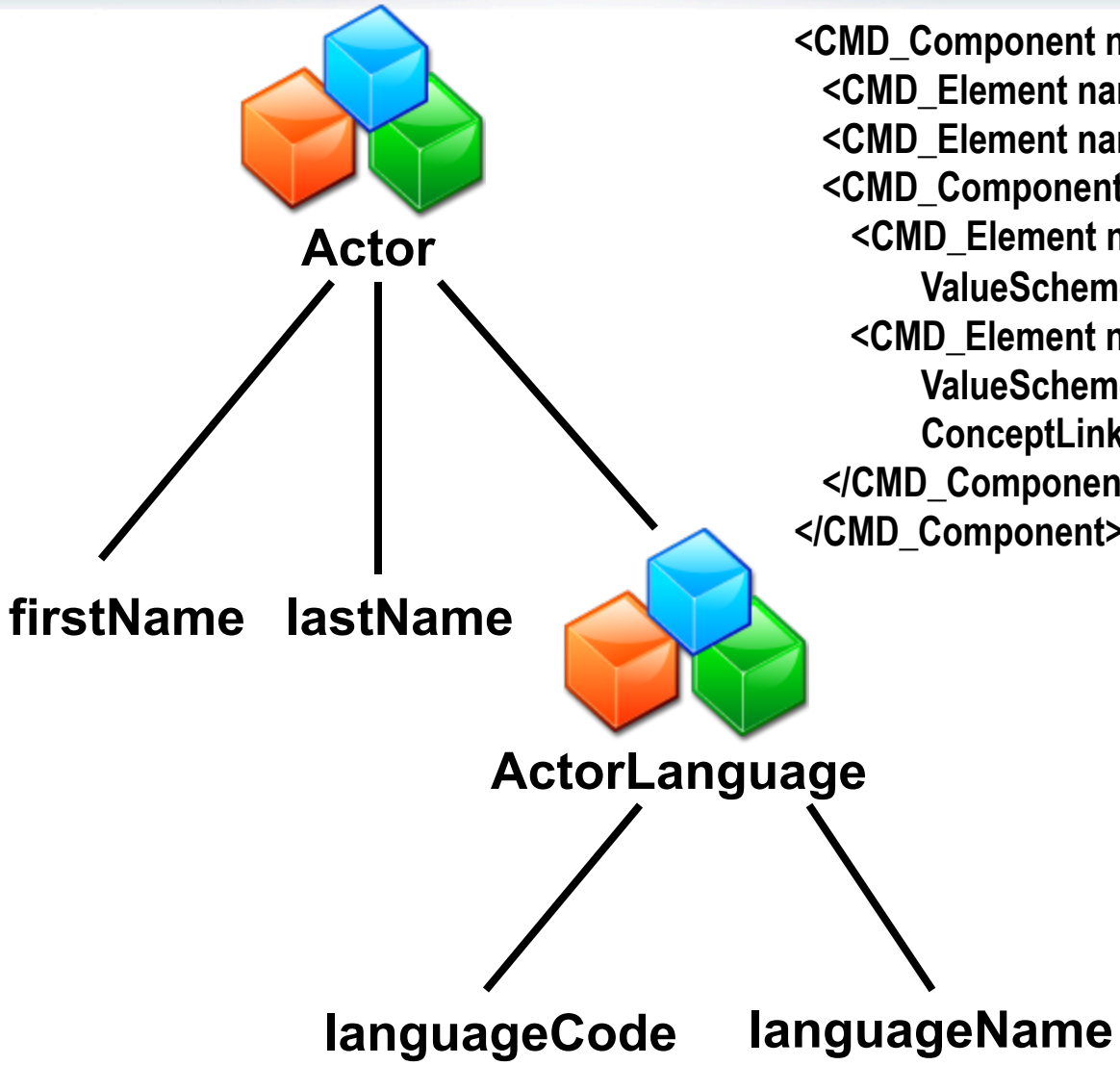
Data Categories



The big picture



Building a component




```
<CMD_Component name="Actor">
  <CMD_Element name="firstName" ValueScheme="string“/>
  <CMD_Element name="lastName" ValueScheme="string"/>
  <CMD_Component name="ActorLanguage">
    <CMD_Element name="LanguageCode“
          ValueScheme="string“ />
    <CMD_Element name="LanguageName“
          ValueScheme="string“
          ConceptLink="http://www.isocat.org/datcat/DC-1766"/>
  </CMD_Component>
</CMD_Component>
```

Using a component

Actor

```
[...]
<Actor>
  <firstName>Louis</firstName>
 <lastName>Couperus</lastName>
 <ActorLanguage>
        <LanguageCode>nld</LanguageCode>
        <LanguageName>Dutch</LanguageName>
 </ActorLanguage>
</Actor>
[...]
```

firstName: Louis

lastName: Couperus

ActorLanguage

languageCode: nld

languageName: Dutch

A close look at a CMDI file

- A toy example:
  - http://www.clarin.eu/cmd/example/example-md-instance.cmdi
- A corpus description:
  - http://www.clarin.eu/cmd/example/example-phonological-corpus.cmdi

Process overview

- Check the Component registry:
  - Any profile that fits your needs?
  - If not:
    - Any component that fits your needs?
    - If not:
      - Create your own component!
      - Looking for a data category that is not there?
        - Create a new data category!
    - Combine components together in a profile
- Start Arbil or your XML-editor
  - Select the profile/XSD that suits your needs
  - Create metadata instances

What is out there?

- About 50 profiles and 250 components: http://catalog.clarin.eu/ds/ComponentRegistry
- About 700 metadata data categories: http://www.isocat.org/
- About 420.000 metadata records:
  - http://catalog.clarin.eu/oai-harvester/
  - http://www.clarin.eu/vlo/
- Planned:
  - Profiles matching the MetaShare schema (XSD)

Conclusions

- You can build your own components and profiles
- You can create metadata descriptions based on a CMDI profile
- What there is now:
  - Component registry
  - CMDI editor: Arbil
  - CMDI facet browser: Virtual Language Observatory
  - Format will be supported in the future
- Planned:
  - Collection Browsers (supporting hierarchies)
  - Sophisticated search engines
  - Repositories
  - Additions to / extensions of the existing software

More information?

- http://www.clarin.eu/cmdi

Where to get support?

- cmdi@clarin.eu

Thank you for your attention

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