CMDI Task Force

Agenda

1. proposal for 1.3 or 1.2.3
   a. recap
   b. Going beyond CLAVAS
   c. vocabularies (open/closed, internal/external)
   d. Translations
   e. 1.3 or 1.2.3?
2. CMDI 1.2 editors
   a. COMEDI (Paul Meurer)
   b. HuC CMDI editor (Rob Zeeman)
3. Lookout to 2.0
4. Future of the CMDI Taskforce

Vocabulary support in CMDI
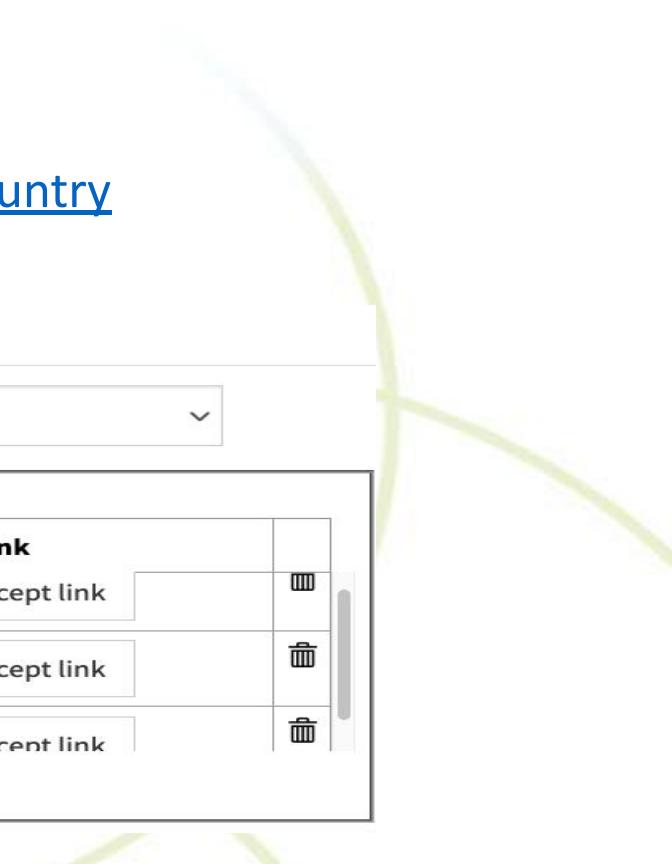
- Pre CMDI 1.2
  - Embedded closed vocabularies, e.g., iso-country
    - Included in the XML schema, i.e., validated!

Vocabulary support in CMDI

- CMDI 1.2
  - Link to CLAVAS vocabulary
    - Closed: embedded + CLAVAS link, e.g., [iso-639-3](iso-639-3)
      - (selection of) CLAVAS vocabulary is copied to embedded closed vocabulary
    - Open: CLAVAS link only
      - Link instance to vocabulary item via ConceptLink

```
<cmdp:org cmd:ValueConceptLink="http://.../clavas/...">Meta</cmdp:org>
```

Vocabulary support in CMDI

CLAVAS

https://vocabularies.clarin.eu/clavas/

- A vocabulary service originally developed in CLARIN-NL
- Provides an API to (incrementaly) search for vocabulary items, to be used by, e.g., an editor
- Contains just one vocabulary
    - ISO 639-3 language codes + labels
        - not the latest, but in sync with components/profiles in use
- On the shelf
    - Licenses
    - MIME types
    - Organisations (needs sync)
- Why so little?
    - Prevent proliferation: one vocab per domain
    - Who claims ownership to deal with change request?
    - How to deal with versioning/subsets?

CMDI & vocabs: going beyond CLAVAS

- CMDI 1.2 always assumes an external vocabulary lives in CLAVAS
    - and will use its OpenSKOS API for interaction
- CLAVAS requires clear ownership of the vocabulary to include it
    - so CLARIN can guarantee persistency & maintenance

- allowing everyone to use their own external vocabulary?
    - common API? SKOSMOS v1
    - persistency guarantees? CLARIN only for CLAVAS (to migrate to SKOSMOS) and (maybe) some other recommendations (e.g. CLARIAH+ FAIR vocab registry)
    - other can be used by your profile/component but **without** CLARIN guarantees!

- CLAVAS will remain strict on ownership
    - who? not the TF, but the SCCTC or …

- versioning strategies are still under discussion, e.g. by SSH Vocab Commons

CMDI 1.2

```
<ValueScheme>
    <Vocabulary
URI="https://hdl.handle.net/11459/CLAVAS_810f8d2a-6723-3ba6-2e57-41d6d3844816"
    ValueProperty="skos:notation">
        <enumeration>
            <item AppInfo="Ghotuo (aaa)"
ConceptLink="http://cdb.iso.org/lg/CDB-00132443-001">aaa</item>
            <item AppInfo="Alumu-Tesu (aab)"
ConceptLink="http://cdb.iso.org/lg/CDB-00133770-001">aab</item>
            <item AppInfo="Ari (aac)"
ConceptLink="http://cdb.iso.org/lg/CDB-00133769-001">aac</item>
            <item AppInfo="Amal (aad)"
ConceptLink="http://cdb.iso.org/lg/CDB-00133768-001">aad</item>
            <item AppInfo="Arbëreshë Albanian (aae)"
ConceptLink="http://cdb.iso.org/lg/CDB-00133767-001">aae</item>
            <item AppInfo="Aranadan (aaf)"
ConceptLink="http://cdb.iso.org/lg/CDB-00133766-001">aaf</item>
            …
        </enumeration>
    </Vocabulary>
</ValueScheme>
```

proposal

```xml
<ValueScheme>
    <Vocabulary
      Endpoint= "https://clarin-skosmos.sd.di.huc.knaw.nl/rest/v1/"
      EndpointType="skosmos-v1"

URI="https://hdl.handle.net/11459/CLAVAS_810f8d2a-6723-3ba6-2e57-41d6d3844816"
      ValueProperty="skos:notation"
      Mode="Closed">
        <Enumeration >
          <Documentation xml:lang="en">...</Documentation>
          <Item ConceptLink="http://cdb.iso.org/lg/CDB-00132443-001">
             <Value>aaa</Value>
              <Label xml:lang="en">Ghotuo (aaa)</Label>
              <Label xml:lang="nl">Afrikaanse taal (aaa)</Label>
          </Item>
          …
        </Enumeration>
    </Vocabulary>
</ValueScheme>
```

Vocabulary types

Three dimensions:

1. **Value domain restriction**
   (open/closed)
2. **Definition inside the component**
   (yes/no)
3. **Reference to external vocabulary**
   (yes/no)
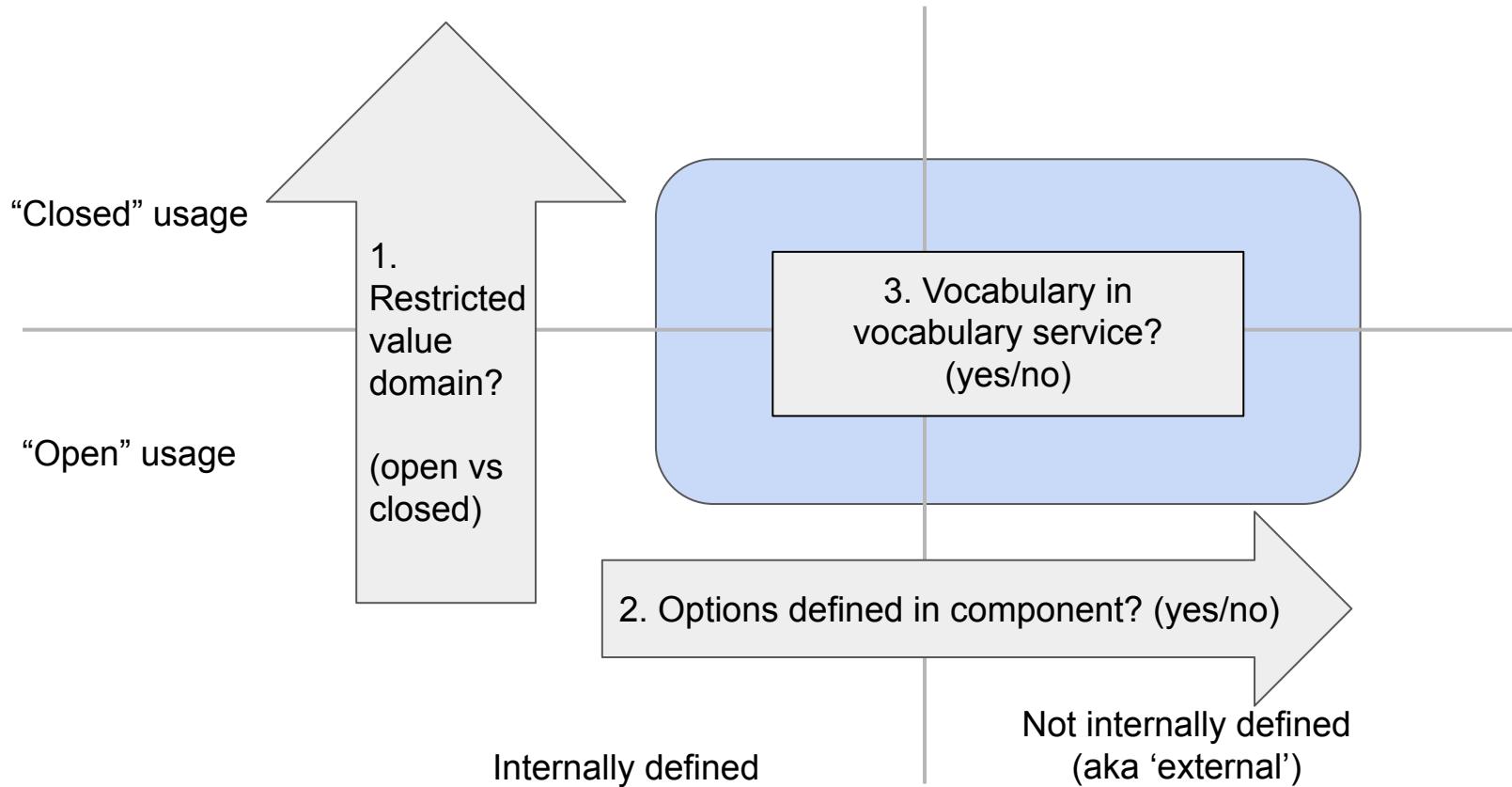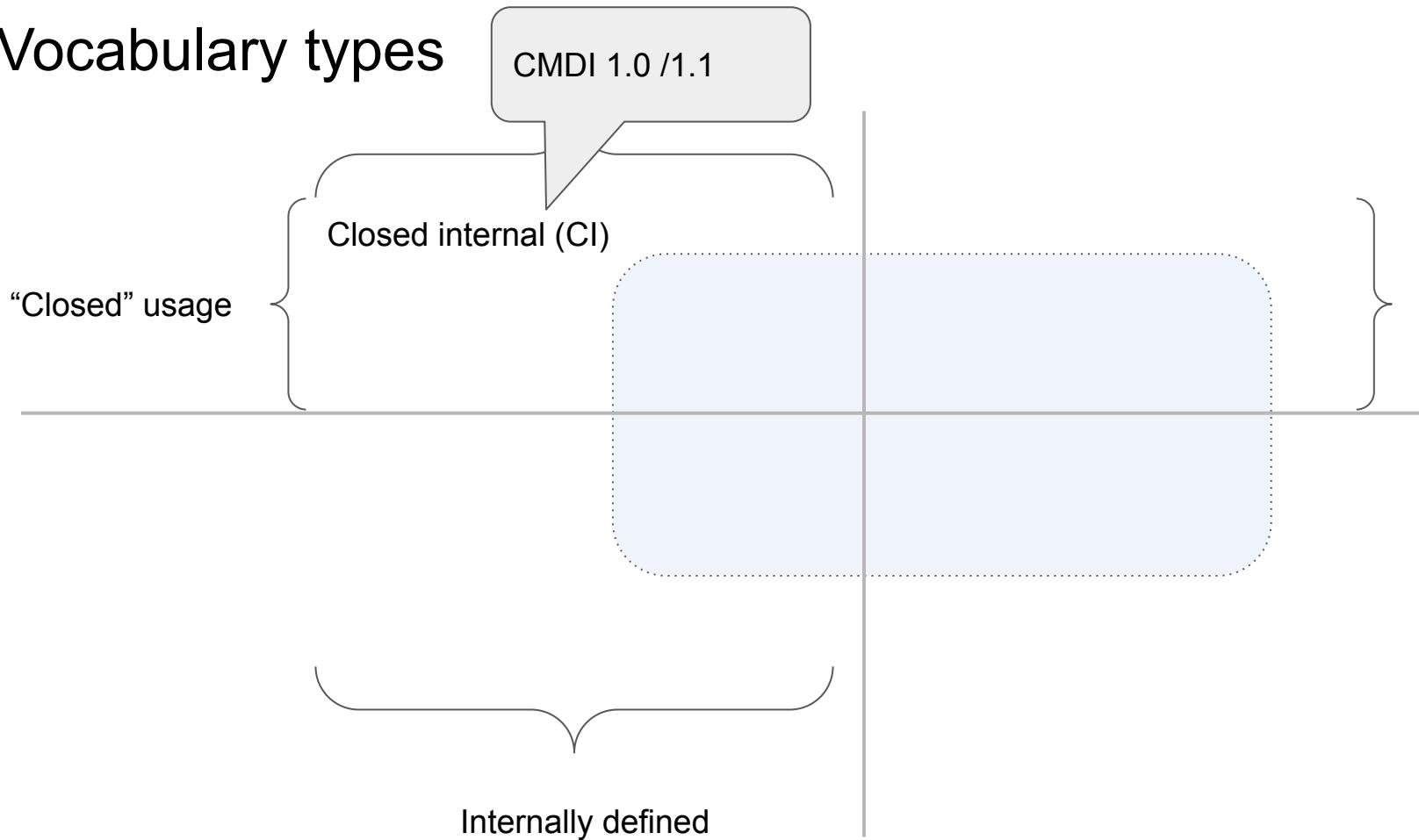
Note:
There are 2 "invalid" combinations, and therefore **6 vocabulary types**!

Vocabulary types

"Closed" usage

"Open" usage

Internally defined

Not internally defined
(aka 'external')

Vocabulary types

"Closed" usage

"Open" usage

1. Restricted value domain?

(open vs closed)

2. Options defined in component? (yes/no)

3. Vocabulary in vocabulary service? (yes/no)

Internally defined

Not internally defined (aka 'external')

Vocabulary types

CMDI 1.0 /1.1

Closed internal (CI)

"Closed" usage
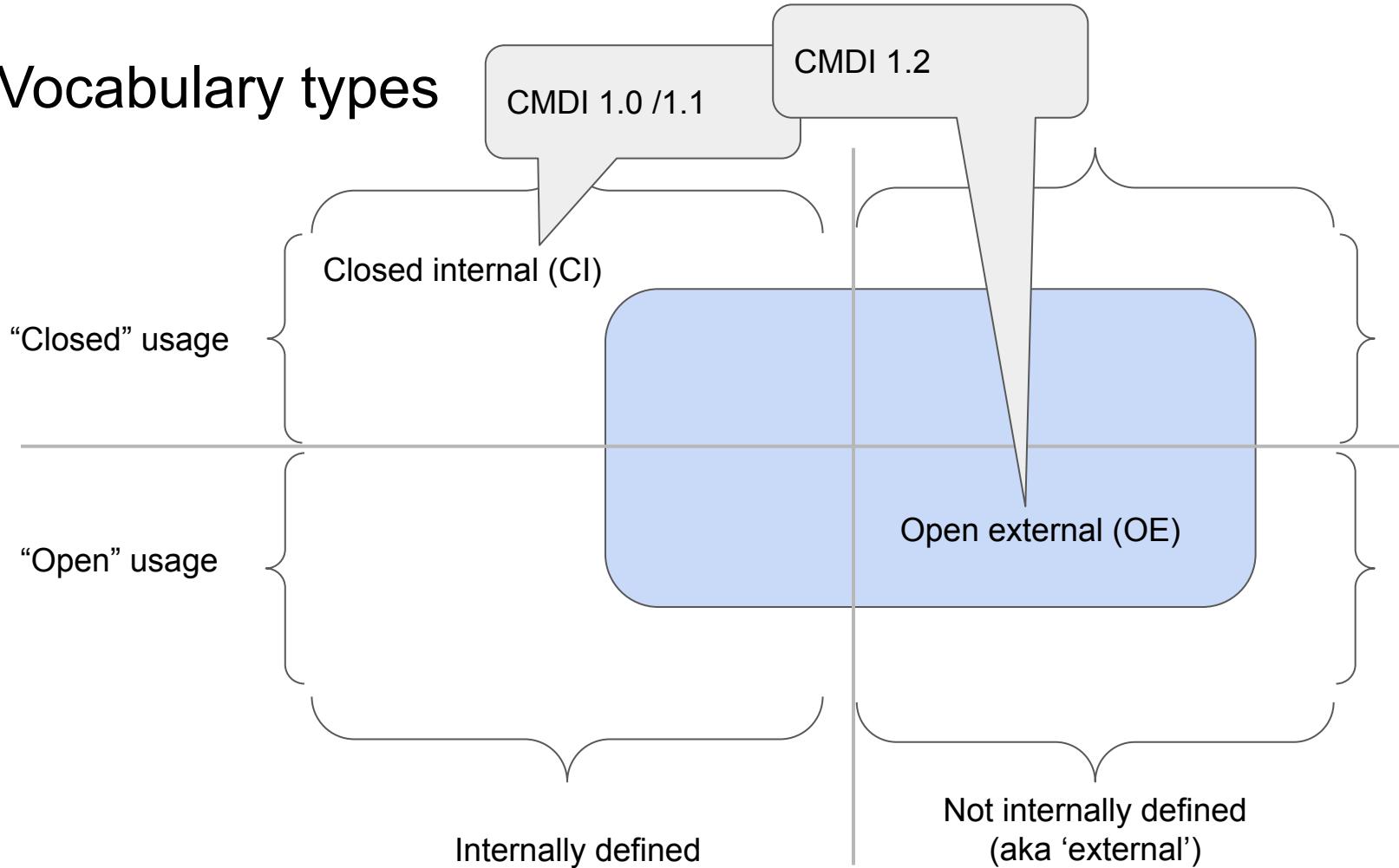
Internally defined
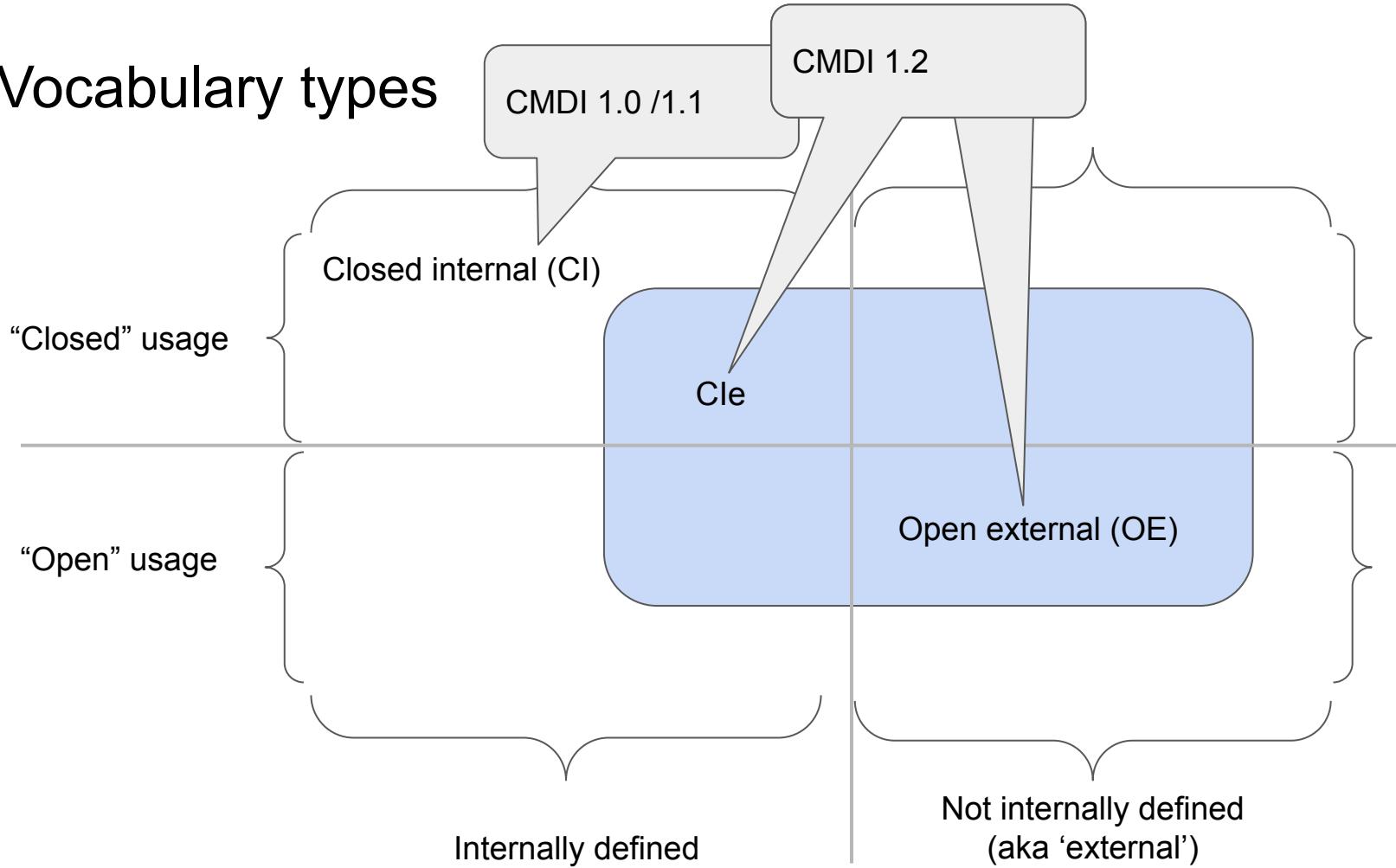
Closed internal vocabularies (CMDI 1.1)

- **Component modeller**: full control over value domain
- **Metadata author**: mandatory choice from predefined options
- **Advantage**: value context is fully known and stable
- **Disadvantage**: change to vocabulary implies a change to the component and downstream changes
- **Use cases**: small, conceptually finite vocabularies
  - binary choices (with or without unspecified, unknown)
  - types (e.g. corpus type, annotation type, global media types)

Vocabulary types

CMDI 1.0 /1.1

CMDI 1.2

Closed internal (CI)

"Closed" usage

Open external (OE)

"Open" usage

Internally defined

Not internally defined
(aka 'external')

Vocabulary types
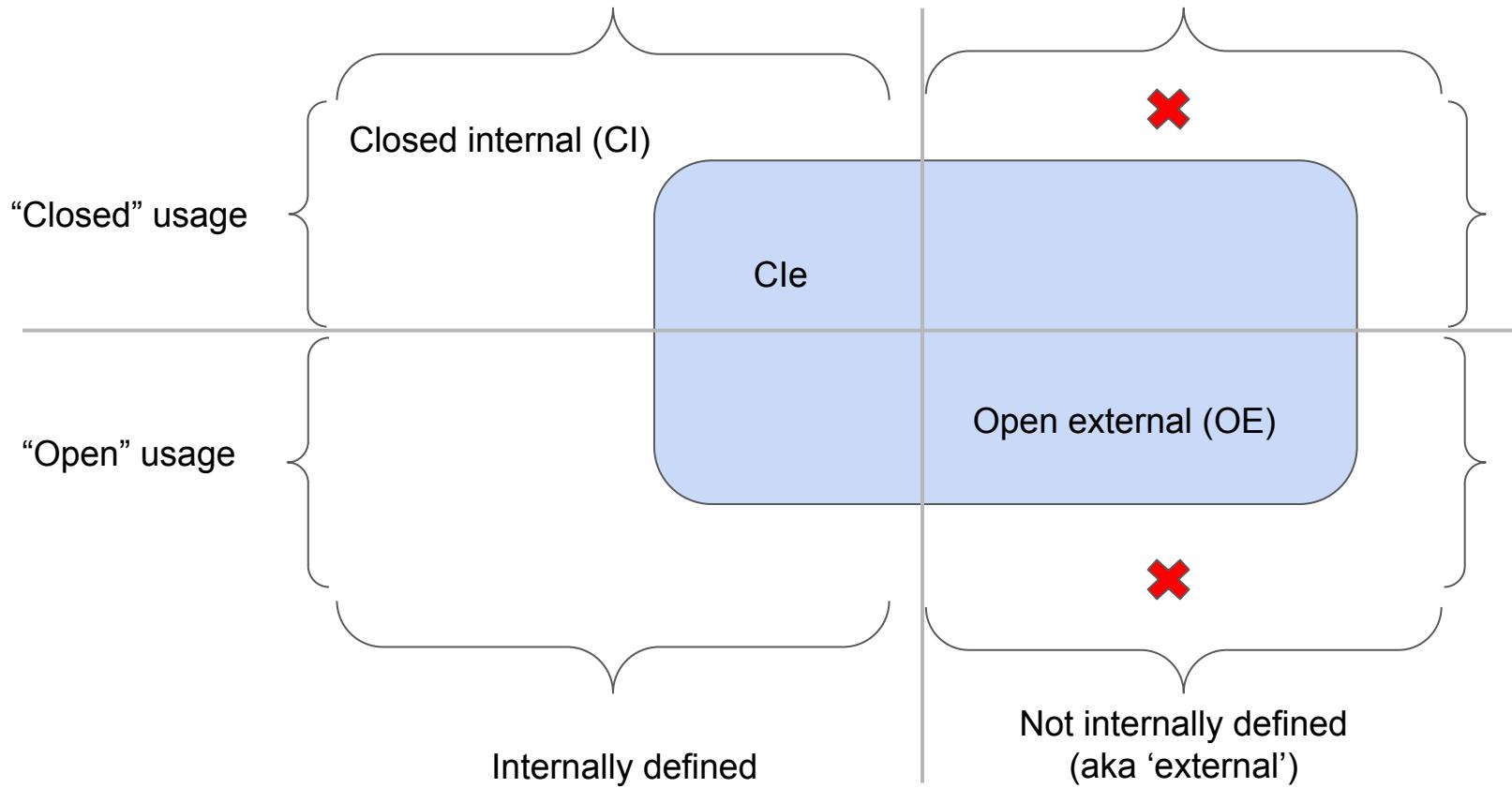
Open external vocabularies (CMDI 1.2)

- **Component modeller:** <u>no</u> full control over value domain
- **Metadata author:** guidance but full control over value
- **Advantage:** flexible value domain for guidance; vocabulary service features
- **Disadvantage:** stability of value domain is not guaranteed
- **Use cases:** large, open vocabularies that may change and may be incomplete
  - Lists of languages, countries etc (change over time)
  - Lists of organisations (expected to be incomplete)
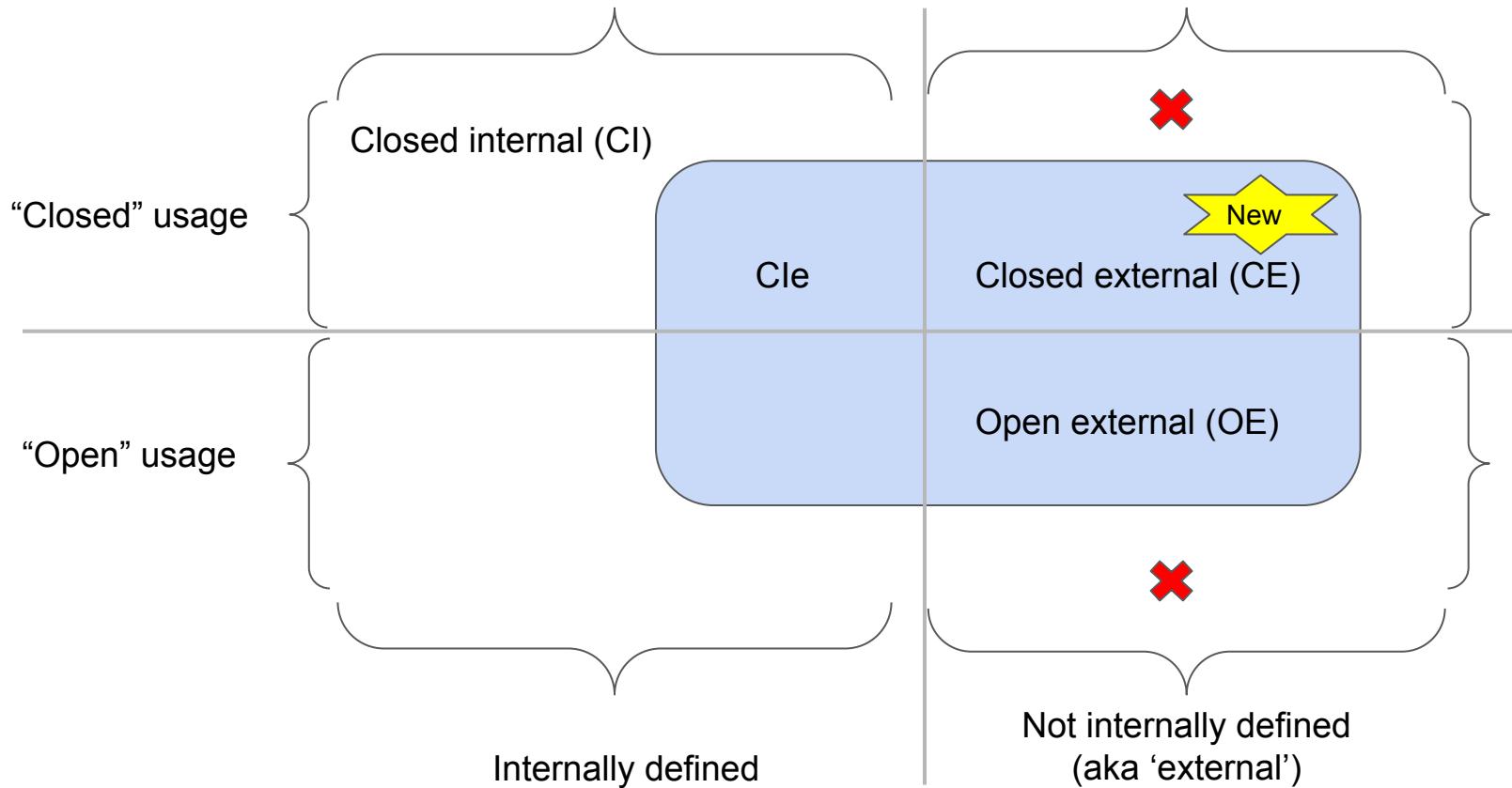
Closed internal with external vocabulary reference (CMDI 1.2)

- **Component modeller:** full control over value domain (like CI), but values are linked to entries from a vocabulary service
- **Metadata author:** mandatory choice from predefined options (like CI)
- **Advantage:** value context is fully known and stable; vocabulary service features can be used uptionally
- **Disadvantages**
  - consolidated values can get out of sync with items in service
  - change to vocabulary implies a change to the component and downstream changes (like CI)
- **Use cases:** custom selection from a larger vocabulary
  - consolidation where context needs to be preserved (frozen snapshot, e.g. for archiving purposes)
    – e.g. selection of relevant countries from a large countries vocabulary (countries where language X is spoken)

Vocabulary types
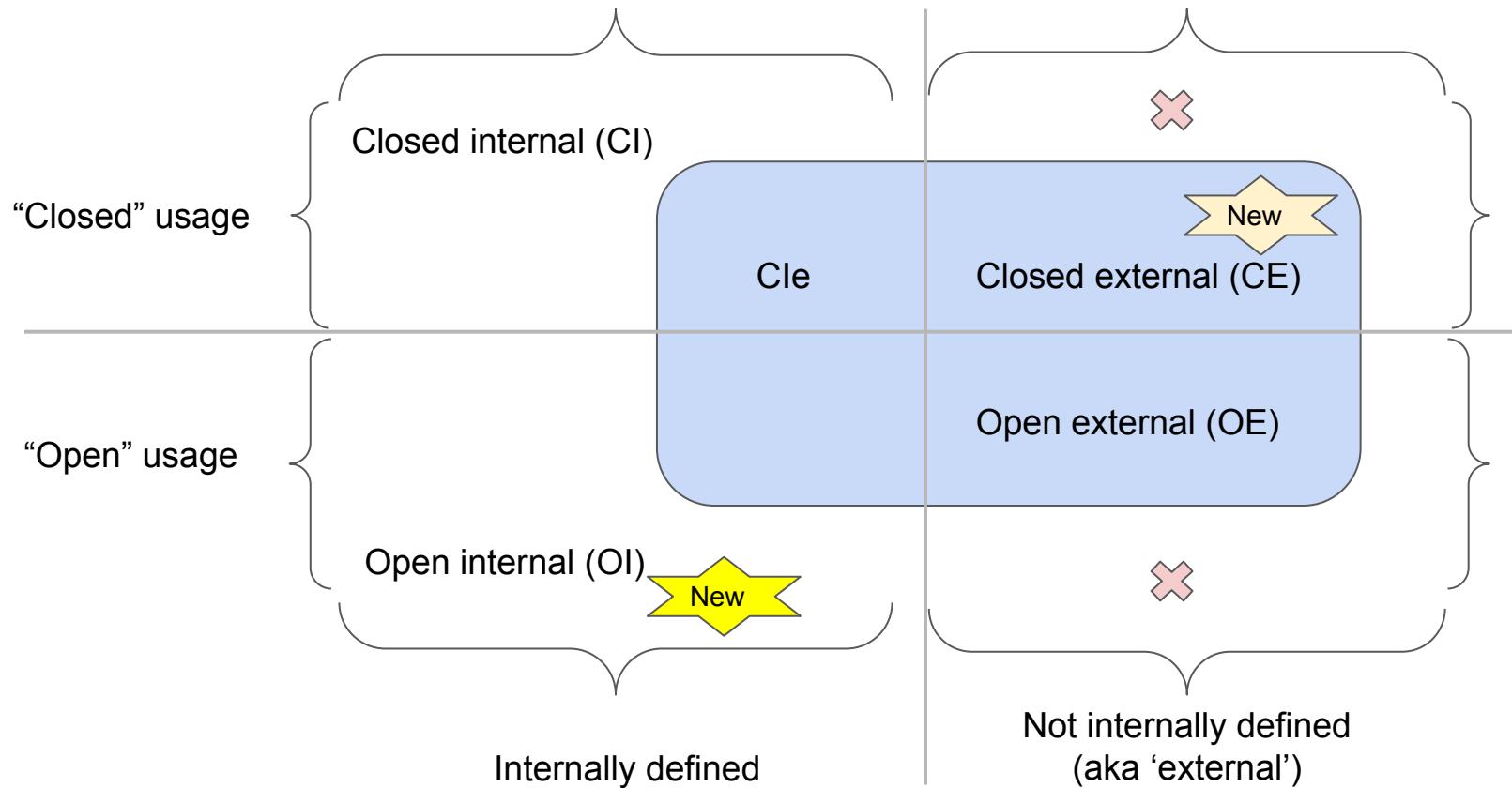
Vocabulary types
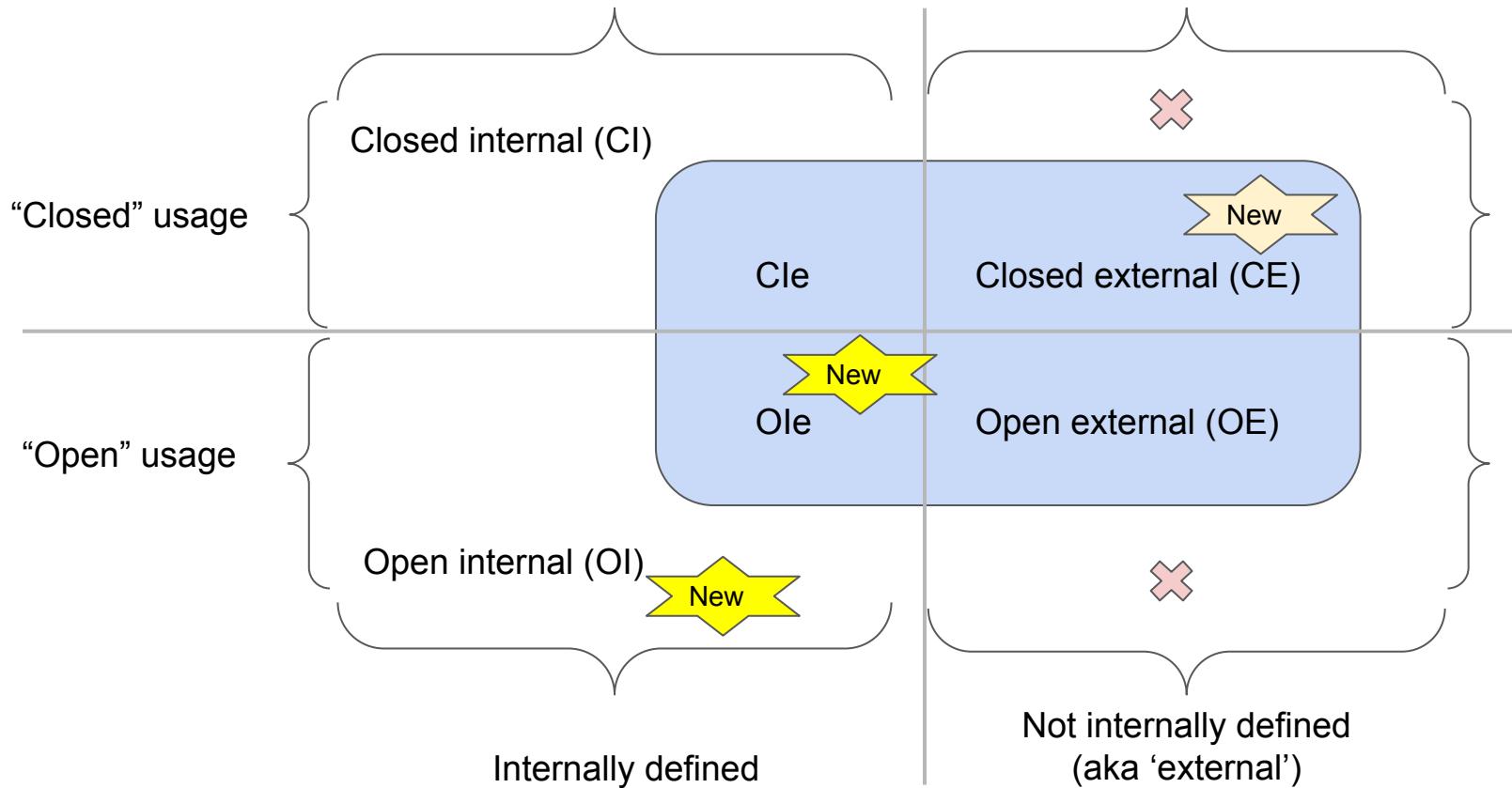
Closed external vocabularies (proposal for CMDI >1.2)

- **Component modeller: exp**resses intention over complete value domain, but no full control
- **Metadata author:** restricted to values of the external vocabulary at time of authoring (not as strictly as closed internal)
- **Advantage:** value domain can be adapted without changing the component definition; vocabulary service features
- **Disadvantage:** stability of value domain is not guaranteed
- **Use cases:** strict usage of vocabularies that are fairly stable but may change occasionally
  - typically broadly used vocabularies that may have high maintenance effort
  - e.g. language codes, subject headings

Vocabulary types

"Closed" usage

Closed internal (CI)

CIe

Closed external (CE)

New

"Open" usage

Open internal (OI)

New

Open external (OE)

Internally defined

Not internally defined
(aka 'external')

Vocabulary types
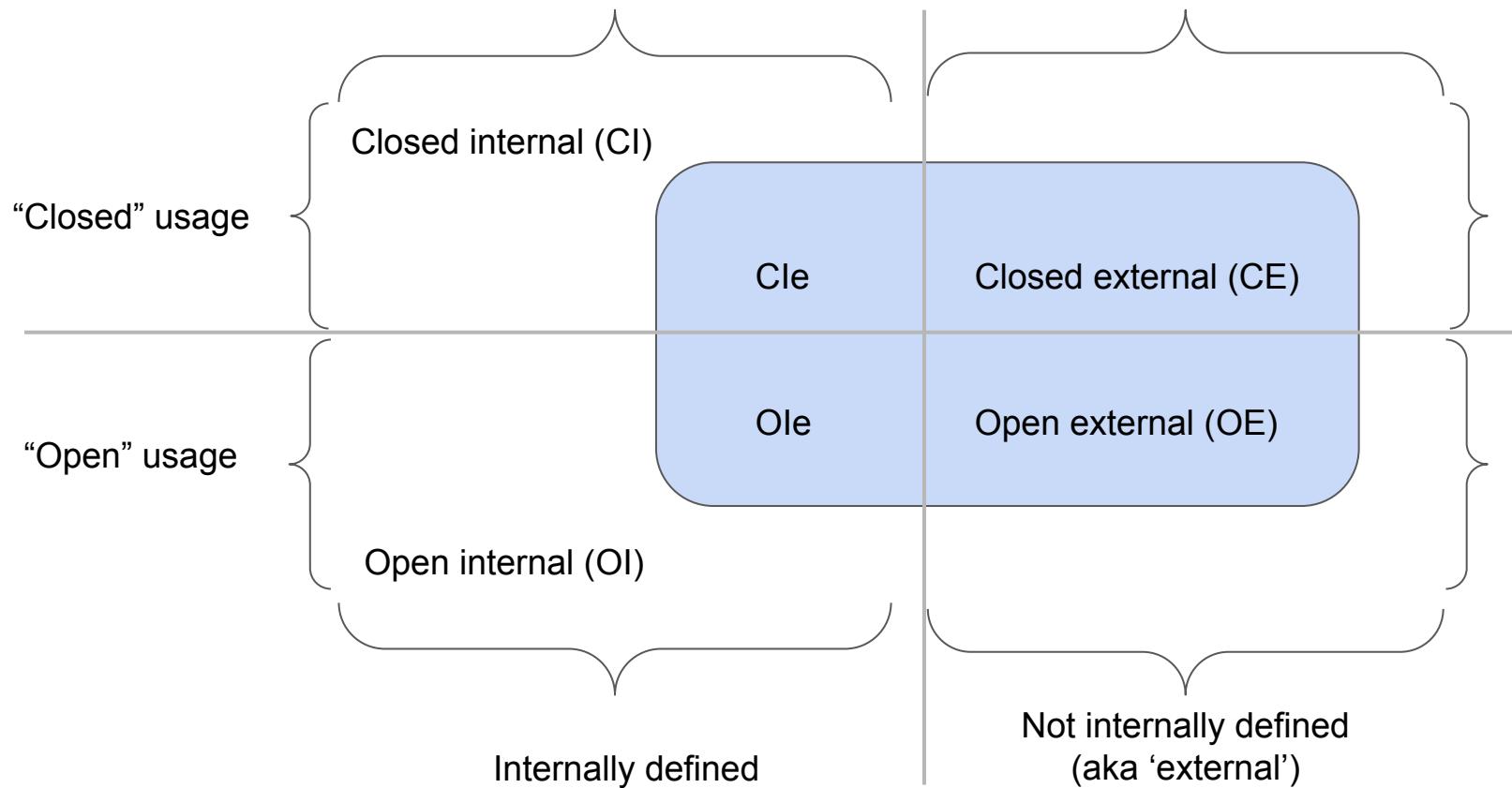
Open internal vocabularies (proposal for CMDI >1.2)
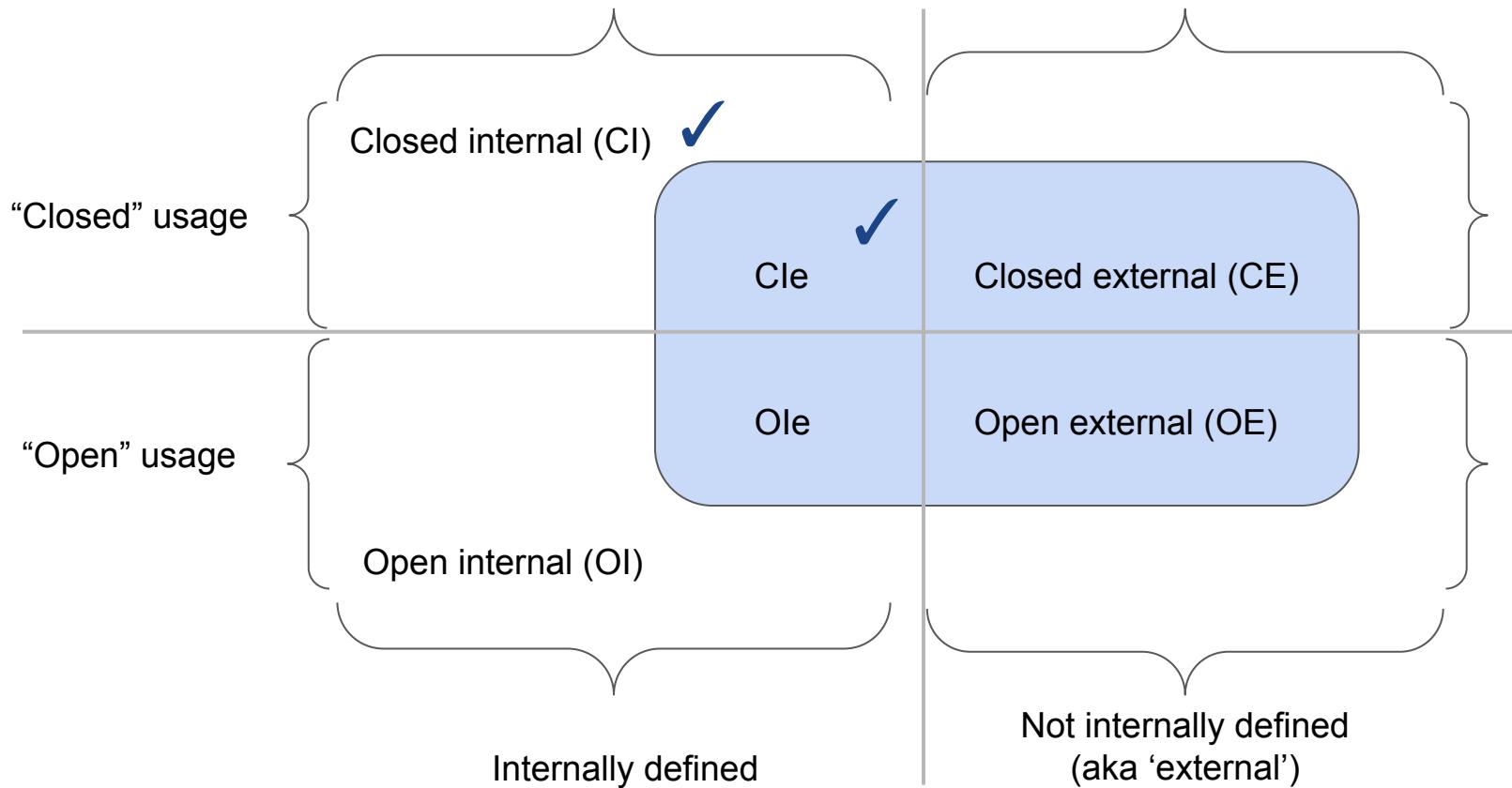
- **Component modeller: only interested in providing value suggestions**
- **Metadata author: guided by predefined values, full control over actual value**
- **Advantage: flexibility for the modeller and author (not bound by an existing vocabulary); value harmony through guidance**
- **Disadvantage: no stable value domain; no vocabulary service features**
- **Use cases:** small relatively ad-hoc value domains for which there are commonly used values that can be used as guidance (gender,) but no external vocabulary is available

External vocabulary reference adds advantage of vocabulary service features

Vocabulary types: overview including proposed types

"Closed" usage

Closed internal (CI)

CIe

Closed external (CE)

"Open" usage

OIe

Open external (OE)

Open internal (OI)

Internally defined

Not internally defined
(aka 'external')
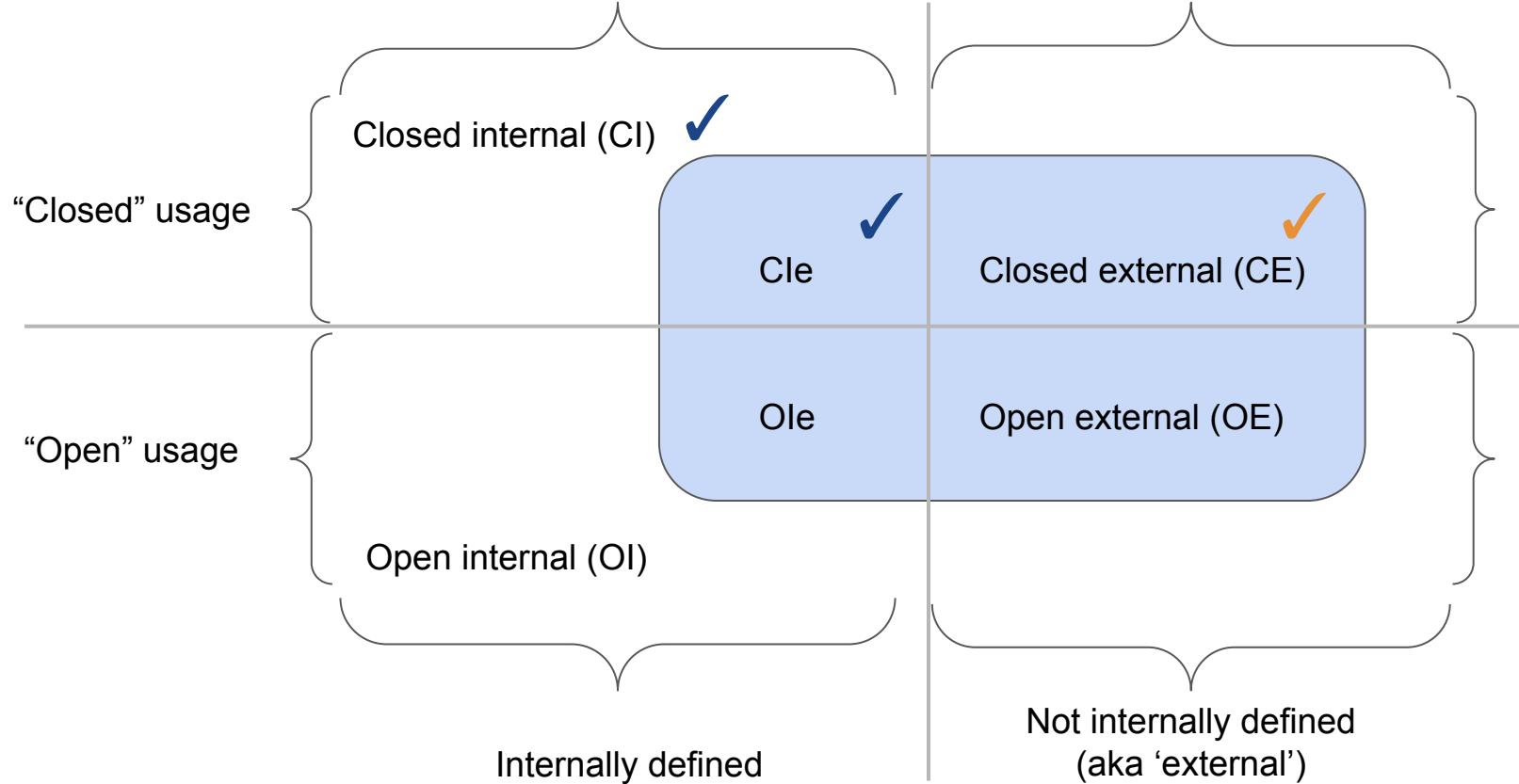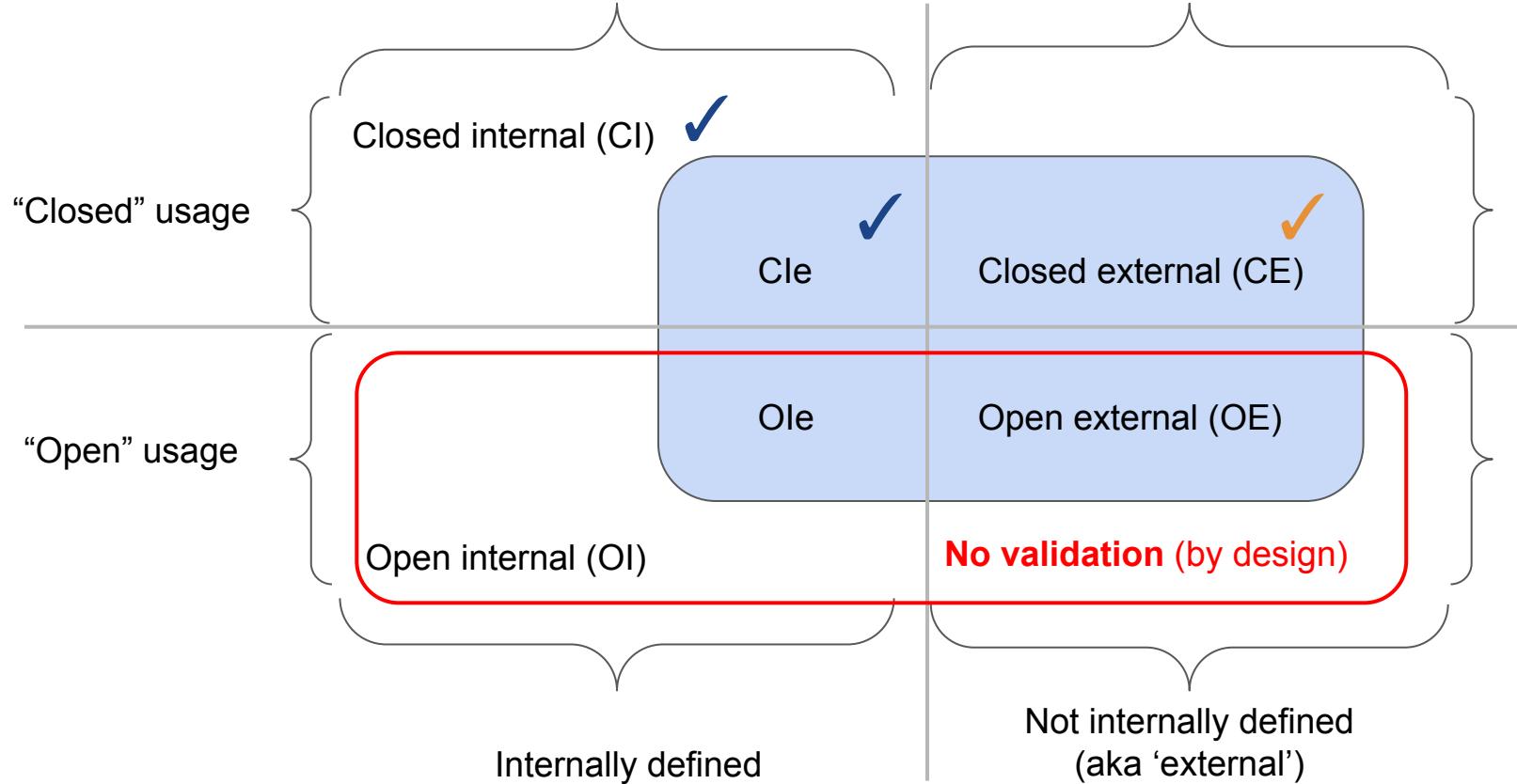
XSD validation of value domain?

XSD validation of value domain?
Schematron validation of value domain?

"Closed" usage

"Open" usage

Closed internal (CI) ✔

CIe ✔   Closed external (CE) ✔

OIe   Open external (OE)

Open internal (OI)

Internally defined

Not internally defined (aka 'external')

XSD validation of value domain?
Schematron validation of value domain?

Closed internal (CI) ✔

"Closed" usage

Cie ✔ Closed external (CE) ✔

"Open" usage

Oie Open external (OE)

Open internal (OI) **No validation** (by design)

Internally defined

Not internally defined
(aka 'external')

Open question: references outside vocabulary scope

- Open vocabularies with reference to external ("OE" and "OIe"):
  Specification **does not restrict value identifier to vocabulary domain**.
  - Metadata **author** can 'mix and match'
  - Metadata **consumer** does not have enough information to process this
  - Best way to deal with this??
    - Document as bad practice?
    - Check with schematron and issue warning?
    - Embrace??

Multilingual labels for components & friends

- editors & viewers currently fallback to the technical names
  - https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu%3Acr1%3Ap_12 97242111880&registrySpace=public
  - https://clarino.uib.no/comedi/editor/foobarMenzo
- if we can add labels to all levels we can create a multilingual rendering:
  - https://menzowindhouwer.github.io/lab/cr2html/index.html#clarin.eu:cr1:p_144042 6460262
  - https://cmdi.sd.di.huc.knaw.nl/ccf/index.php?page=metadata&id=2
  - https://github.com/knaw-huc/clariah-cmdi-parser/blob/master/examples/tweaks/M eertensCollectionTweak.xml
- We could use a cue?
  - cue:cmd_label_nl="Auteur"
- Proposal: neat multilingual labels in the profile/component spec

```xml
<ValueScheme>
    <Vocabulary
      Endpoint= "https://clarin-skosmos.sd.di.huc.knaw.nl/rest/v1/"
      EndpointType="skosmos-v1"

URI="https://hdl.handle.net/11459/CLAVAS_810f8d2a-6723-3ba6-2e57-41d6d3844816"
      ValueProperty="skos:notation"
      Mode="Closed">
        <Enumeration >
          <Documentation xml:lang="en">...</Documentation>
          <Item ConceptLink="http://cdb.iso.org/lg/CDB-00132443-001">
            <Value>aaa</Value>
            <Label xml:lang="en">Ghotuo (aaa)</Label>
            <Label xml:lang="nl">Afrikaanse taal (aaa)</Label>
          </Item>
          …
        </Enumeration>
    </Vocabulary>
</ValueScheme>
```

CMDI 1.3 or 1.2.3?

1. 1.3 has impact on the record
   a. but the proposed changes don't impact the record!

```
<cmd:CMD xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:cmd="http://www.clarin.eu/cmd/1"
  xmlns:cmdp="http://www.clarin.eu/cmd/1/profiles/clarin.eu:cr1:p_1288172614026"
CMDVersion="1.2"
  xsi:schemaLocation="http://www.clarin.eu/cmd/1
https://infra.clarin.eu/CMDI/1.x/xsd/cmd-envelop.xsd
http://www.clarin.eu/cmd/1/profiles/clarin.eu:cr1:p_1288172614026
https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p
_1288172614026/xsd">
```

CMDI 1.3 or 1.2.3?

1. 1.2.3 no impact on any processing
   a. fixing some XSD inspired tags & adding translations in the profiles does impact tools!

CMDI 1.2 editors

1. COMEDI presentation
2. HuC CMDI editor presentation

Lookout to 2.0

- keep profiles in the Component Registry, but support different serializations of records
    - tools use the profile as-is (maybe serialized differently
    - profile translated into serialization specific validation schemas
- XML
    - tools shouldn't rely on a specific XSD transformation, instead use the profile as-is
    - validate using profile to XSD
    - attributes are maybe too XML specific and hard to express in other record serializations

        <cmdp:weight unit="kg" certainty="0.7">76</cmdp:weight>

- RDF
    - validate using profile to SHACL (?)
    - problems:
        - turning nesting into predicates
        - no attributes on elements
            - maybe via RDF*?
- JSON
    - validate using profile to JSON schema (?)
    - problems:
        - attributes on elements (?)

Future of the CMDI task force(s)

-