CLARIN
Common Language Resources and Technology Infrastructure

Max Planck Institute for Psycholinguistics

INL
INSTITUUT VOOR NEDERLANDSE LEXICOLOGIE

Meertens instituut

CLARIN-NL
Metadataproject

Griet Depoorter

griet.depoorter@inl.nl

Instituut voor Nederlandse Lexicologie

19/02/2010

Projectpartners

- Daan Broeder
- Dieter Van Uytvanck

- Folkert de Vriend

- Laura van Eerten
- Griet Depoorter

Structuur presentatie

- Traditionele en componentmetadata
- Componentmetadata binnen CLARIN
- CMD-componenten: eigenschappen, principes, praktijk en problemen
- XML-toolkit en ISOcat DC registry: werking en ervaringen

Componentmetadata (1)

- Traditionele metadata
  - Veel verschillende standaarden
  - Niet flexibel
  - Niet onderling uitwisselbaar
  - Vaak beperkt tot / gericht op specifieke soort resource

Componentmetadata (2)

- Componentmetadata
  - Flexibiliteit: gebruiker kiest / maakt zelf componenten
  - Geschikt voor verschillende soorten resources
  - Conceptlinks naar datacategorieën (ISOcat data category registry) en relation registry
  - Andere standaarden kunnen uitgedrukt worden in componentmetadata

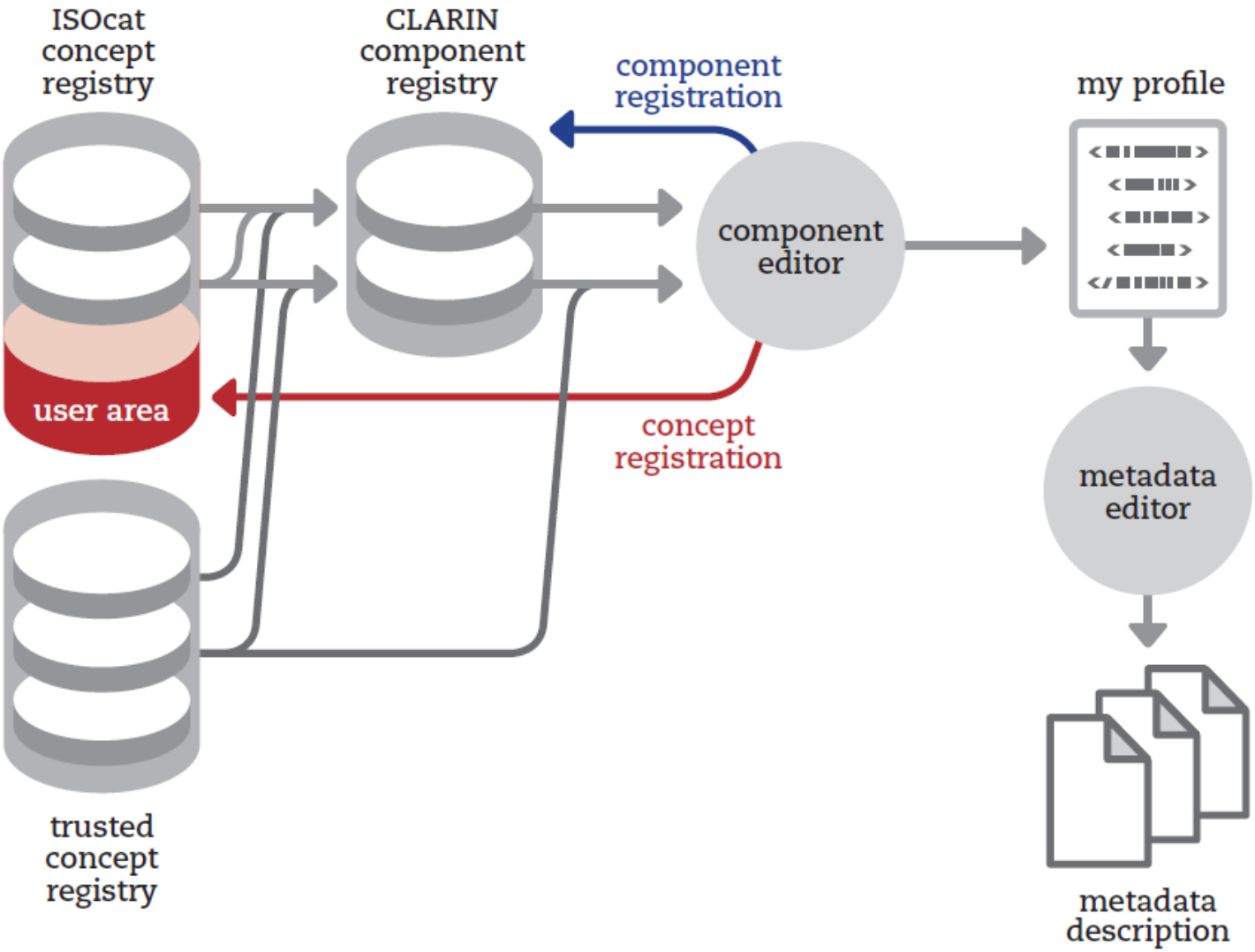
Componentmetadata (3)

- Terminologie:
  - **Element** = kerneenheid (een “veld”) – bv. *Age*
  - **Component** = een verzameling van 1 of meerdere elementen (en componenten) – bv. *Speaker*
  - **Profiel** = een verzameling van componenten – bv. een lexiconprofiel
  - **Schema** = formele grammatica die een profiel beschrijft – bv. *olac.xsd*
  - **Instantie** = een metadatabeschrijving – bv. *PAROLE.xml*

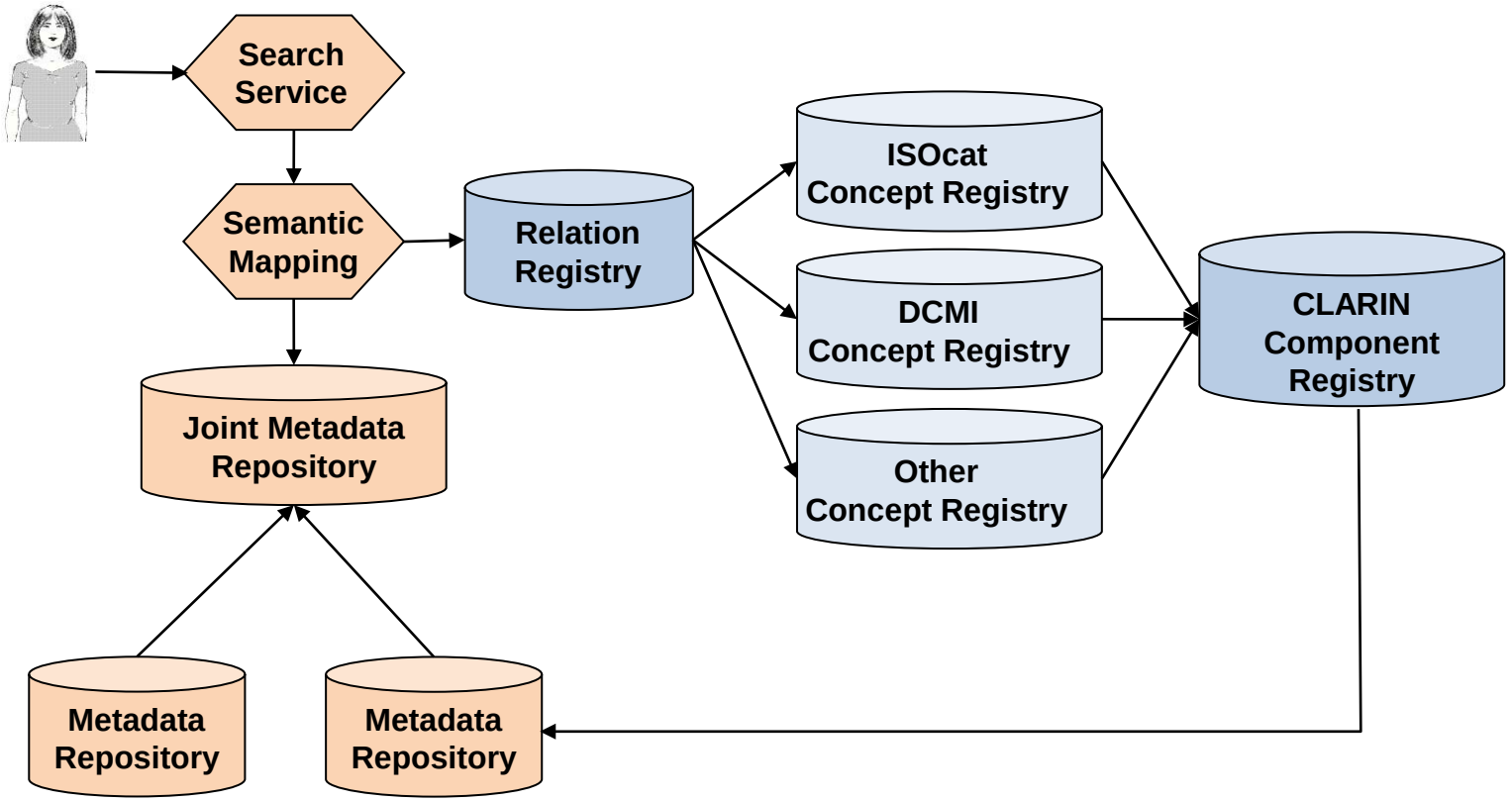
Componentmetadata (4)

```
<CMD_Component name="Speaker_Language" CardinalityMin="0" CardinalityMax="unbounded">
     <CMD_Element name="Mother_Tongue" CardinalityMin="0" CardinalityMax="1">
        <ValueScheme>
            <enumeration>
                <item>Unspecified</item>
                <item>Unknown</item>
                <item>true</item>
                <item>false</item>
            </enumeration>
        </ValueScheme>
    </CMD_Element>
    <CMD_Element name="Primary_Language" ConceptLink="http://www.isocat.org/datcat/DC-2553" CardinalityMin="0" CardinalityMax="1">
        <ValueScheme>
            <enumeration>
                <item>Unspecified</item>
                <item>Unknown</item>
                <item>true</item>
                <item>false</item>
            </enumeration>
        </ValueScheme>
    </CMD_Element>
    <CMD_Component filename="http://www.clarin.eu/cmd/components/clarin-nl/common/component-description.xml" CardinalityMin="0" />
    <CMD_Component filename="http://www.clarin.eu/cmd/components/clarin/cmdi-language.xml" CardinalityMin="1" CardinalityMax="1"/>
</CMD_Component>
```

Componentmetadata binnen CLARIN (1)



Componentmetadata binnen CLARIN (2)



(MPI, Austrian Academy, Språkbanken Univ. Gothenburg, DFKI, IDS)

CMD-componenten (1)

- Hergebruik stimuleren

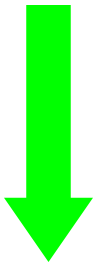
  ➔ project, locatie, taal,…

```
- <Header>
    <ID>eu.clarin.cmdi.language</ID>
    <Name>Language</Name>
    <Description>Component for describing a certain language (free name, ISO-639-3 code)</Description>
  </Header>
- <CMD_Component name="language">
    <CMD_Element name="languageName" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2484" />
    <CMD_Component filename="iso-639-3.xml" CardinalityMin="0" />
  </CMD_Component>
</CMD_ComponentSpec>
```

CMD-componenten (2)

- Hiërarchisch wat de inhoud betreft

  Generieke metadata, van toepassing op breed spectrum resources

  Specifieke metadata, van toepassing op bepaald soort resource

CMD-componenten (3)

Collections

Collectietype, naam, versie, project…

Corpus

(Hoeveelheid) talen, validatie,…

Database

Dimensies (sociale fenomenen, tijd, ruimte),…

Speech Corpus

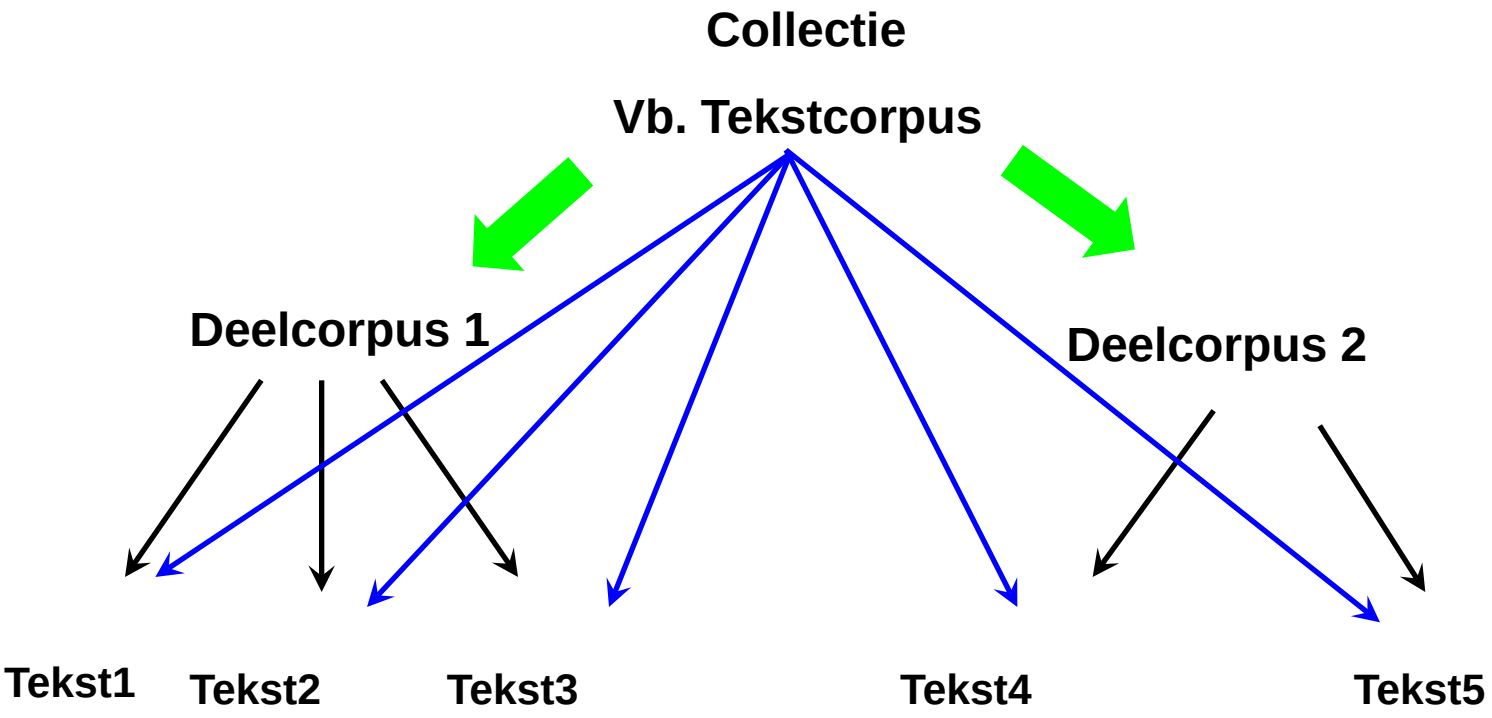
Annotaties, aantal sprekers, duur van de spraak…

Text Corpus

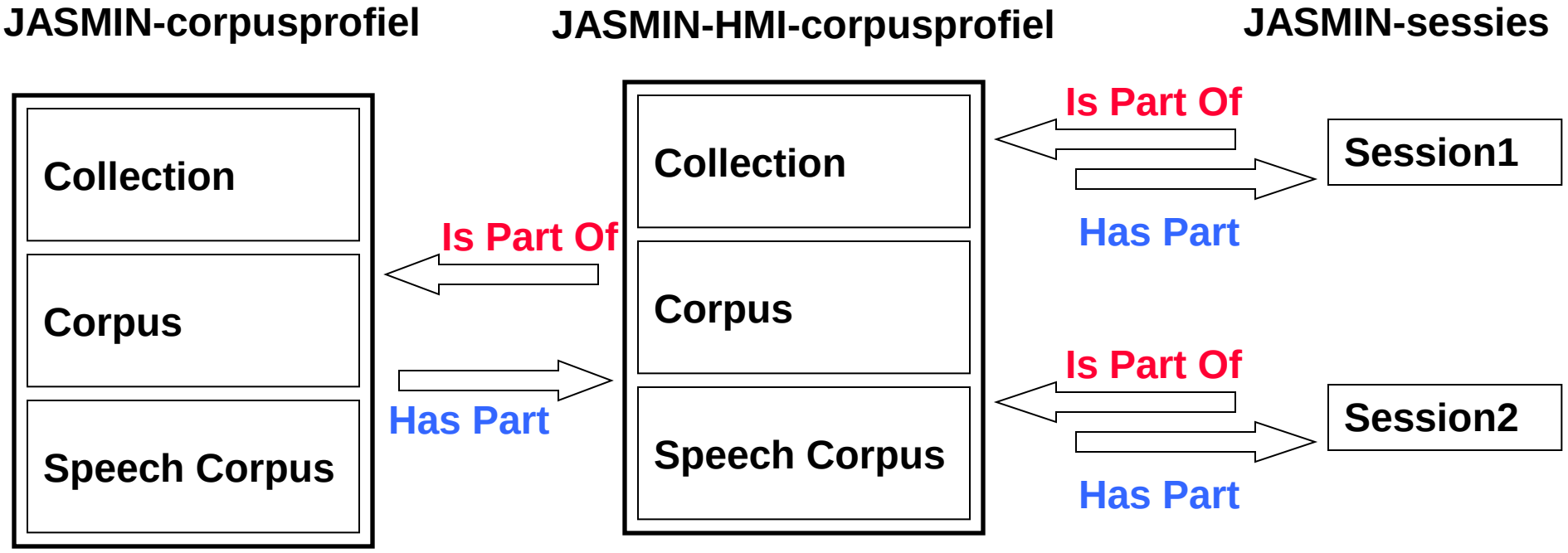
Character encoding, oorsprong teksten,…

CMD-componenten (5)

- Hiërarchisch opgebouwd (granulariteit)



CMD-componenten (6)



CMD-componenten (7)

- Gebaseerd op bestaande metadata (DC, IMDI, OLAC)
- Ongeveer 65 componenten (lexica/woordenboeken, spraakcorpora, tekstcorpora, databanken)
- Componenten worden aangepast n.a.v. concrete metadata-instanties

CMD-componenten - problemen (1)

- Granulariteit: wanneer maak je deelcollecties aan?
  - Als de onderdelen voldoende onderscheidende kenmerken hebben
  - Als de deelresource een geheel vormt en zinvol is
  - Als de resourcecreator wil dat een deelresource apart gevonden/geciteerd moet kunnen worden

CMD-componenten - problemen (2)

- Projectspecifieke codes (vb. leeftijds-categorieën):

  ➔ *projectspecifieke componenten*

- Onderscheid data – metadata kan problematisch zijn

  ➔ *vb. Boedelbank*

CMD-componenten - problemen (3)

- Duplicatie van metadata-informatie

  ➔ vb. meertalig Dutch Parallel Corpus: taalinformatie op corpusniveau en op tekstniveau

XML-toolkit - werking (1)

```
<CMD_Component name="Speaker_Language" CardinalityMin="0" CardinalityMax="unbounded">
     <CMD_Element name="Mother_Tongue" CardinalityMin="0" CardinalityMax="1">
        <ValueScheme>
            <enumeration>
                <item>Unspecified</item>
                <item>Unknown</item>
                <item>true</item>
                <item>false</item>
            </enumeration>
        </ValueScheme>
    </CMD_Element>
    <CMD_Element name="Primary_Language" ConceptLink="http://www.isocat.org/datcat/DC-2553" CardinalityMin="0" CardinalityMax="1">
        <ValueScheme>
            <enumeration>
                <item>Unspecified</item>
                <item>Unknown</item>
                <item>true</item>
                <item>false</item>
            </enumeration>
        </ValueScheme>
    </CMD_Element>
    <CMD_Component filename="http://www.clarin.eu/cmd/components/clarin-nl/common/component-description.xml" CardinalityMin="0" />
    <CMD_Component filename="http://www.clarin.eu/cmd/components/clarin/cmdi-language.xml" CardinalityMin="1" CardinalityMax="1"/>
</CMD_Component>
```

XML-toolkit - werking (2)

XML-toolkit - werking (3)

```
- <Components>
  - <Speaker_Language>
      <Mother_Tongue>true</Mother_Tongue>
      <Primary_Language>true</Primary_Language>
      <description />
    - <language>
        <languageName>Dutch</languageName>
      - <ISO639>
          <iso-639-3-code>dum</iso-639-3-code>
        </ISO639>
      </language>
    </Speaker_Language>
  </Components>
```

XML-toolkit - ervaringen

- Minimale kennis van XML vereist
- Niet echt gebruikersvriendelijk door technische stappen
- Maar: goede XML-editor kan ongemakken opvangen

ISOcat DC registry – werking

- Metadata: 217 datacategorieën beschikbaar
- Doorzoekbaar
- DC’s creëren:
  - private workspace
  - toegang voor beperkte groep
  - public en standaardisatieproces

ISOcat DC registry – ervaringen (1)

- Nieuwe datacategorieën:

  bv. *Legal Owner* en *Mother Tongue*

- Dubbele entry's:

  bv. *Source: A complete citation of the bibliographic information pertaining to a document or other resource.*

  http://www.isocat.org/datcat/DC-1968 en DC-471

- Naam data category = definitie

  bv. *Unknown* en *Unspecified*

  (http://www.isocat.org/datcat/DC-2591 en DC-2592)

ISOcat DC registry – ervaringen (2)

- Inconsequente definities: vb. Contactinformatie

  - DC-2512: The name of the person who was participating in the creation project.
  - DC-2454: The name of the person that can be contacted to get access to the resource or to the tool/service.
  - DC-2505: The address of an organization that was/is involved in creating, managing and accessing resource or tool/service.
  - DC-2521: The email address of a person or an organization that is involved in creating, managing or accessing resources or tools/services.
  - DC-2459: The organization that was leading the creation project or that is responsible for accessing the resource and the contact person is affiliated with.
  - DC-2461: The telephone number of a person or an organization that is involved in creating, managing or accessing the resource.

Ten slotte

- ISOcat data category registry: http://www.isocat.org
- CLARIN-NL-componenten: http://www.clarin.eu/cmd/components/clarin-nl/
- XML-toolkit: http://www.clarin.eu/toolkit
- Best Practicesdocument

Einde