META-SHARE overview

Gunn Inger Lyse (UiB)

Elina Desipri, Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis
(ILSP/RC Athena)

Workshop on the Interoperability of Metadata, Oslo, June 5, 2012

Co-funded by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the contracts T4ME, CESAR, METANET4U, META-NORD (grant agreements no. 249119, 271022, 270893, 270899).

Overview

- Introduction
- META-SHARE – basic features
- Applying META-SHARE: META-NORD user cases in Norway
- Observed challenges
- Conclusion

Overview

- Introduction
- META-SHARE – basic features
- Applying META-SHARE – META-NORD user cases in Norway
- Observed challenges
- Conclusion

META-SHARE (1)

- **META-SHARE main goals:**
  - Establish an open linguistic infrastructure for R&D communities developing language technology.
  - Upgrade, harmonize, document and catalogue language resources and tools.
- **Organization**
  - Cooperation between META-NORD, T4ME, CESAR and METANET4U in the common META-NET network;
  - Funded by the EU under FP7 and CIP.

META-SHARE (2)

- An open and interoperable infrastructure to support R&D in Language Technology:
  - a **network of repositories** of LRT
  - a common set of **metadata** describing LRT
  - central inventories allow for **uniform search and access** to LRT
  - LRT can have both **open and restricted** access rights, **free and for-a-fee**

Overview

- Background
- META-SHARE – basic features
- Applying META-SHARE – META-NORD user cases in Norway
- Observed challenges
- Conclusion

Role of the model

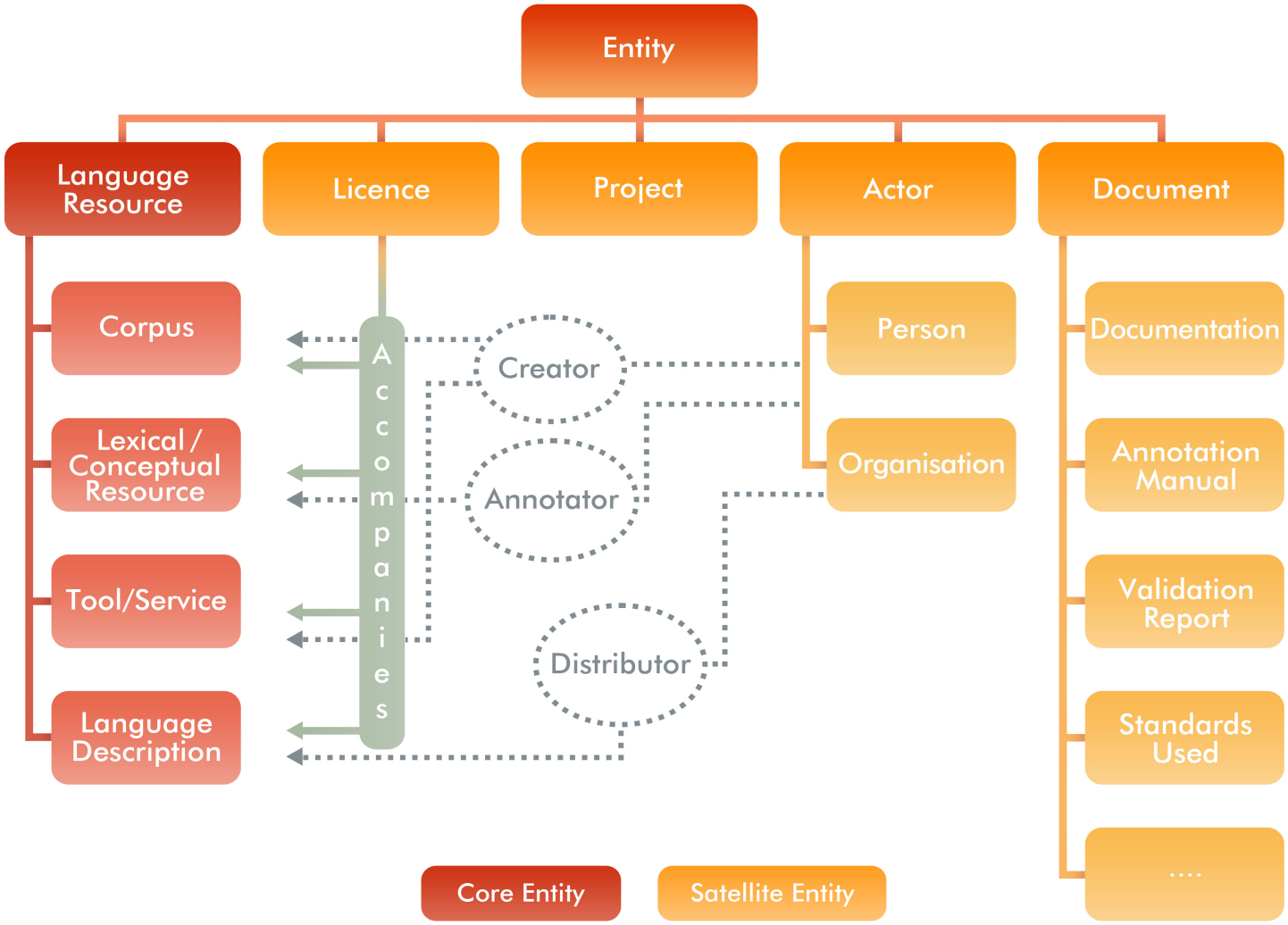
- The metadata descriptions constitute the means by which
  - LR producers describe their resources

  and
  - LR users identify the resources they seek

Design principles

- **expressiveness**: the proposed LR typology aims at covering any type of resource
- **interoperability**: avoid reinventing the wheel by
  - harmonisation of existing schemas and related initiatives
  - harmonisation of licensing templates to ensure legal interoperability
  - extensibility: the modularity of the schema allows for future extensions
- **flexibility**: a minimal and a maximal schema caters for exhaustive but also for minimal descriptions

Ontology


Entity
Language Resource
Licence
Project
Actor
Document
Corpus
Lexical / Conceptual Resource
Tool/Service
Language Description
Accompanies
Creator
Annotator
Distributor
Person
Organisation
Documentation
Annotation Manual
Validation Report
Standards Used
....
Core Entity
Satellite Entity

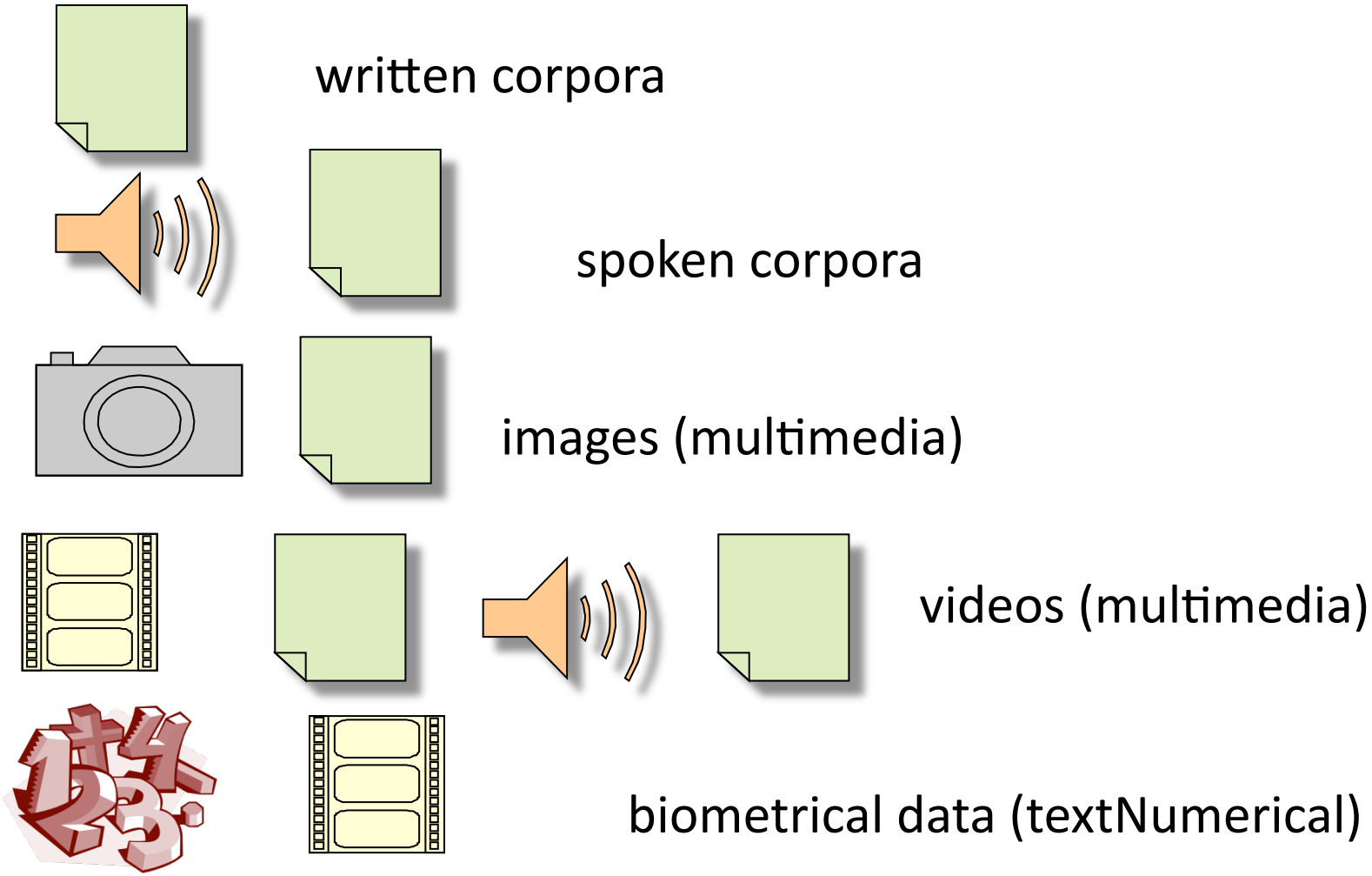
LRs typology (1)

- Two main classification axes:
  - **resourceType**
  - **mediaType**:

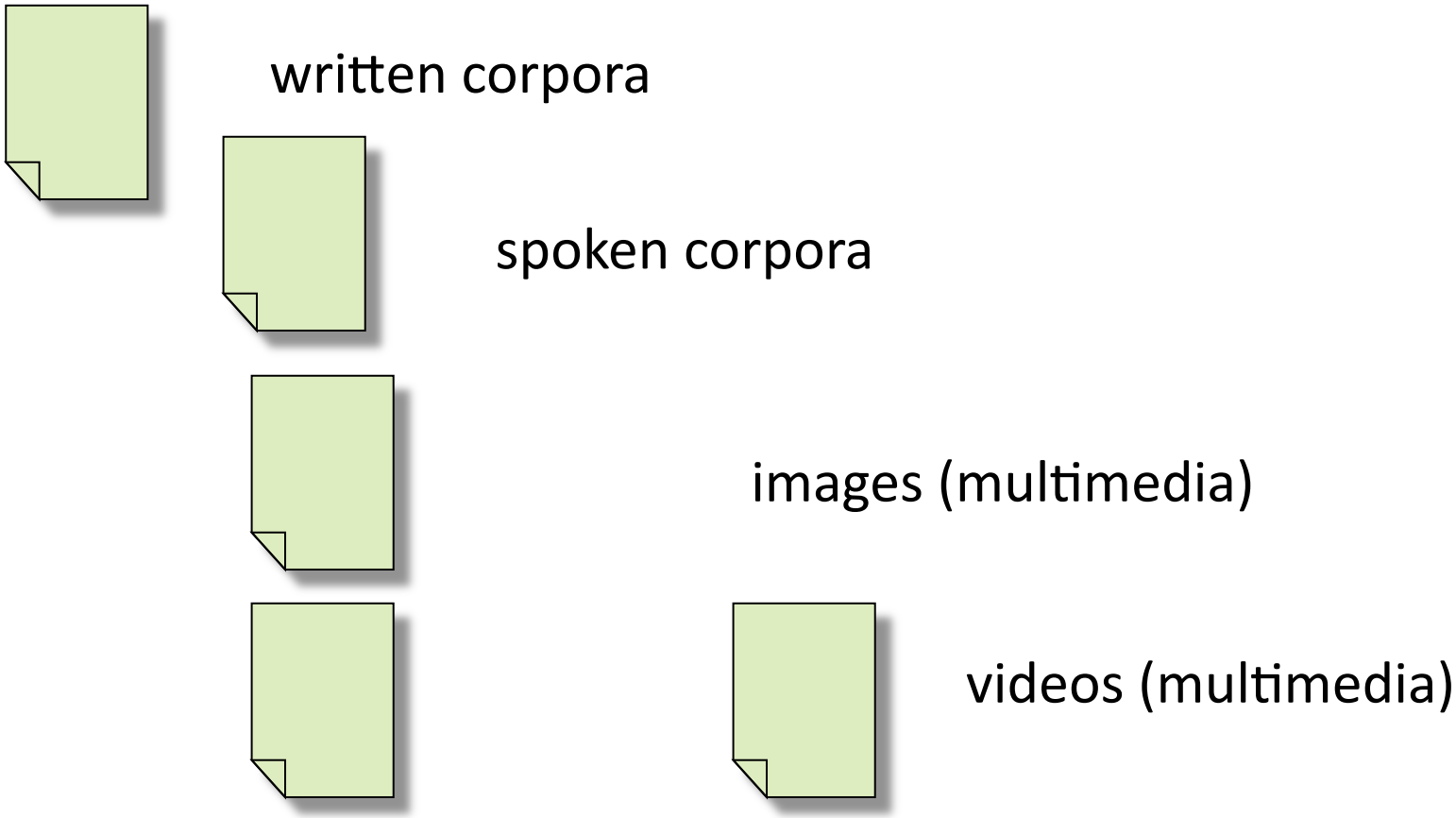
LRs typology (2)

- Two main classification axes:
  - **resourceType**
    - *corpus* (written/text, oral/spoken, multimodal/multimedia corpora)
    - *lexical/conceptual resource* (terminological resources, word lists, semantic lexica, ontologies, etc.,)
    - *tool/service* (processing tools, applications, web services, etc. required for processing data resources)
    - *language description* (grammars, typological databases, courseware, etc.,)
  - **mediaType**: text (+textNumerical and textNgram), audio, image, video

mediaType combinations



Identity intact



META-SHARE structure

- **Profile -** a profile is the set of all the components describing a specific LR type (or subtype)
- **Component -** groups together semantically coherent elements, relations (and other components)
- **Element -** encodes specific descriptive features of the LRs.
- **Relations -** encode linked features between resources.

Minimal & maximal schema

- In order to accommodate flexibility, elements belong to two basic levels of description:
  - **Minimal schema** - minimum set of obligatory elements and relations required for effective LR search and identification
  - **Maximal schema** - additional set of recommended and optional elements and relations for the whole lifecycle of LR production and usage

Minimal & maximal schema

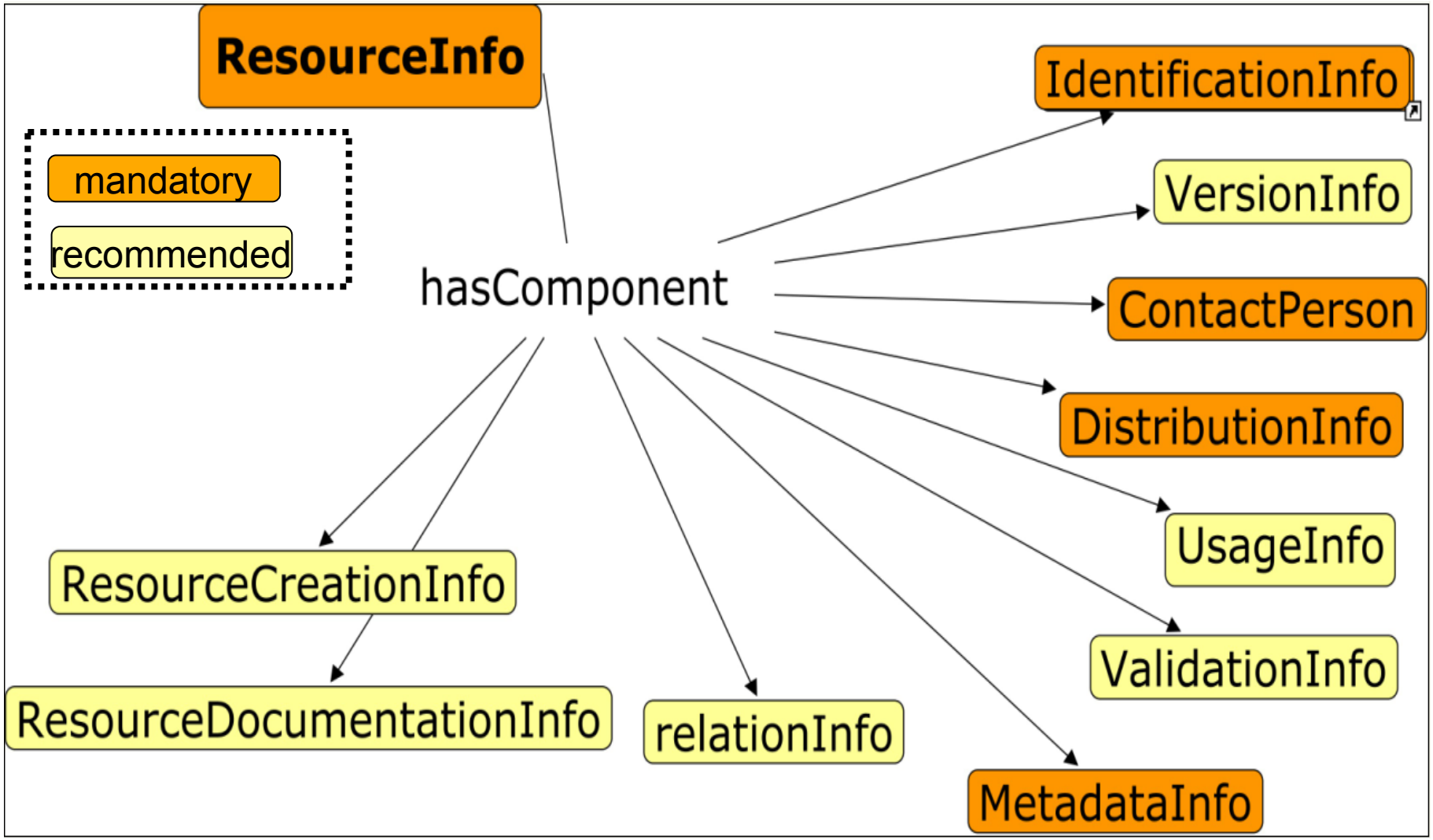
- ***Minimal schema***
  - identification information (title, unique identifier),
  - distribution information, and, if available, licensing details
  - contact person
  - Metadata creation details
- ***Maximal schema***
  - provenance information,
  - creation details,
  - validation/evaluation information,
  - intended and actual use(s) etc.

Mandatory / recommended / optional distinction

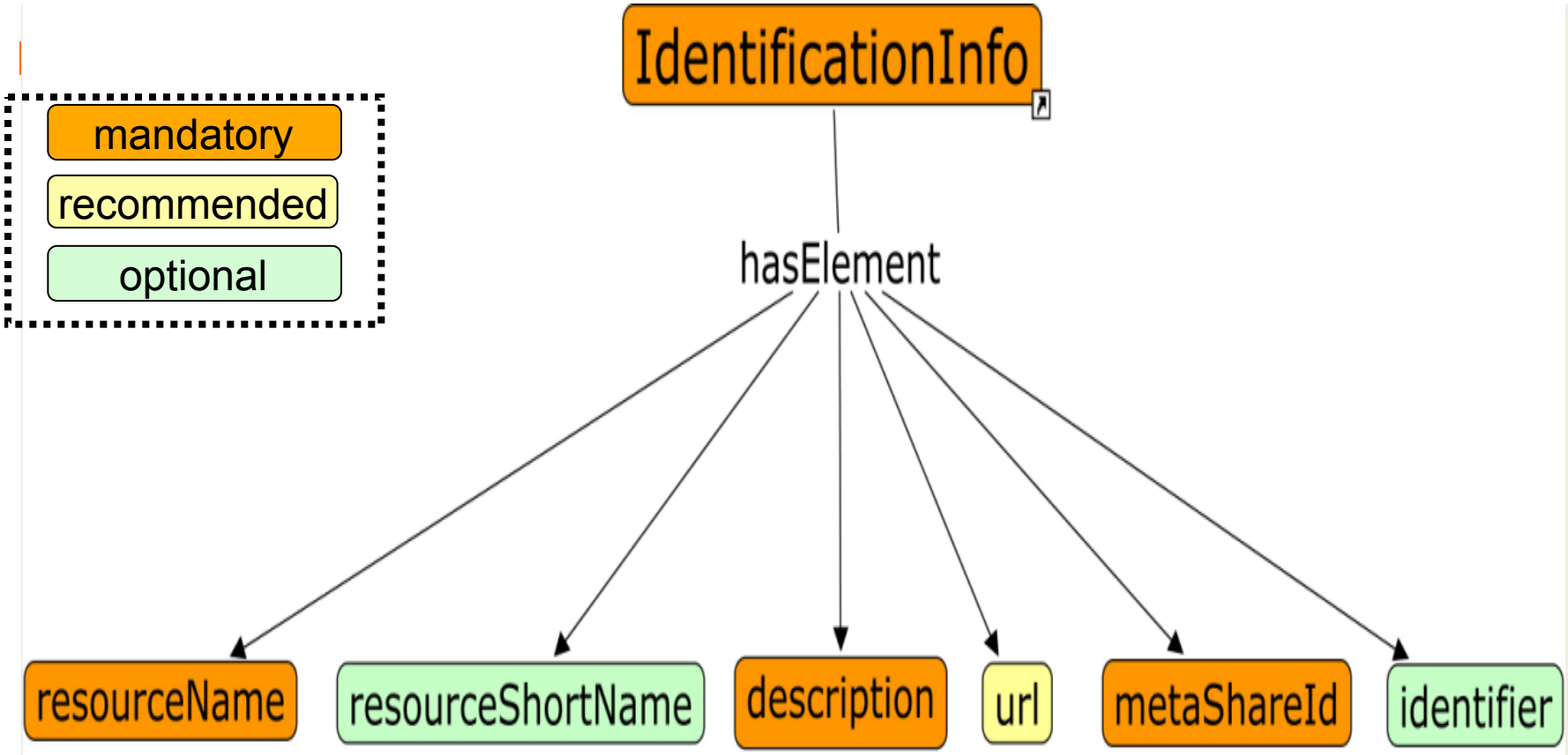
- mandatory components
  - when you describe a LR mandatory components must be used

- mandatory elements
  - mandatory elements inside Recommended or Optional components: if this component is used, then this element is mandatory

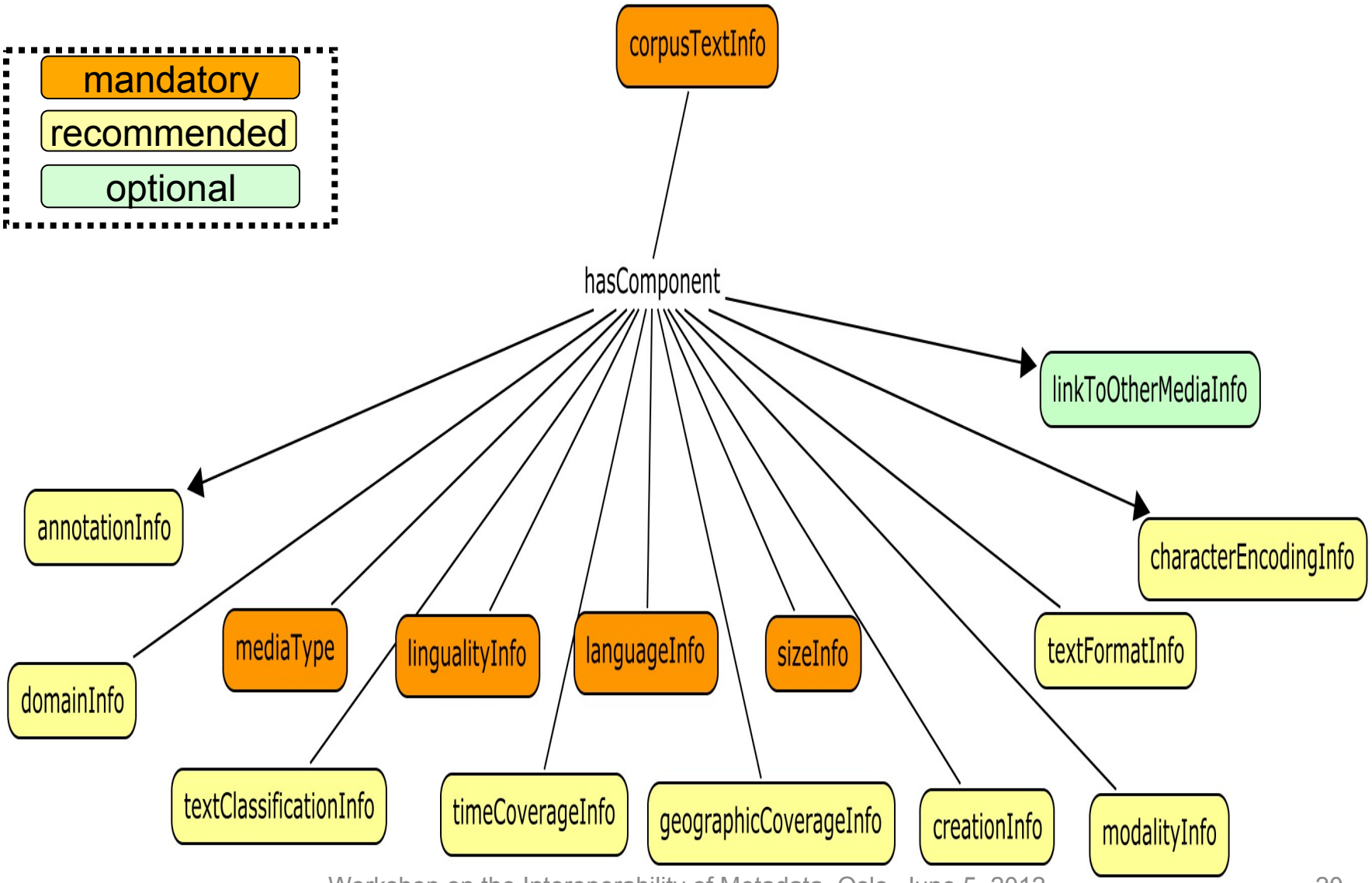
The core description component



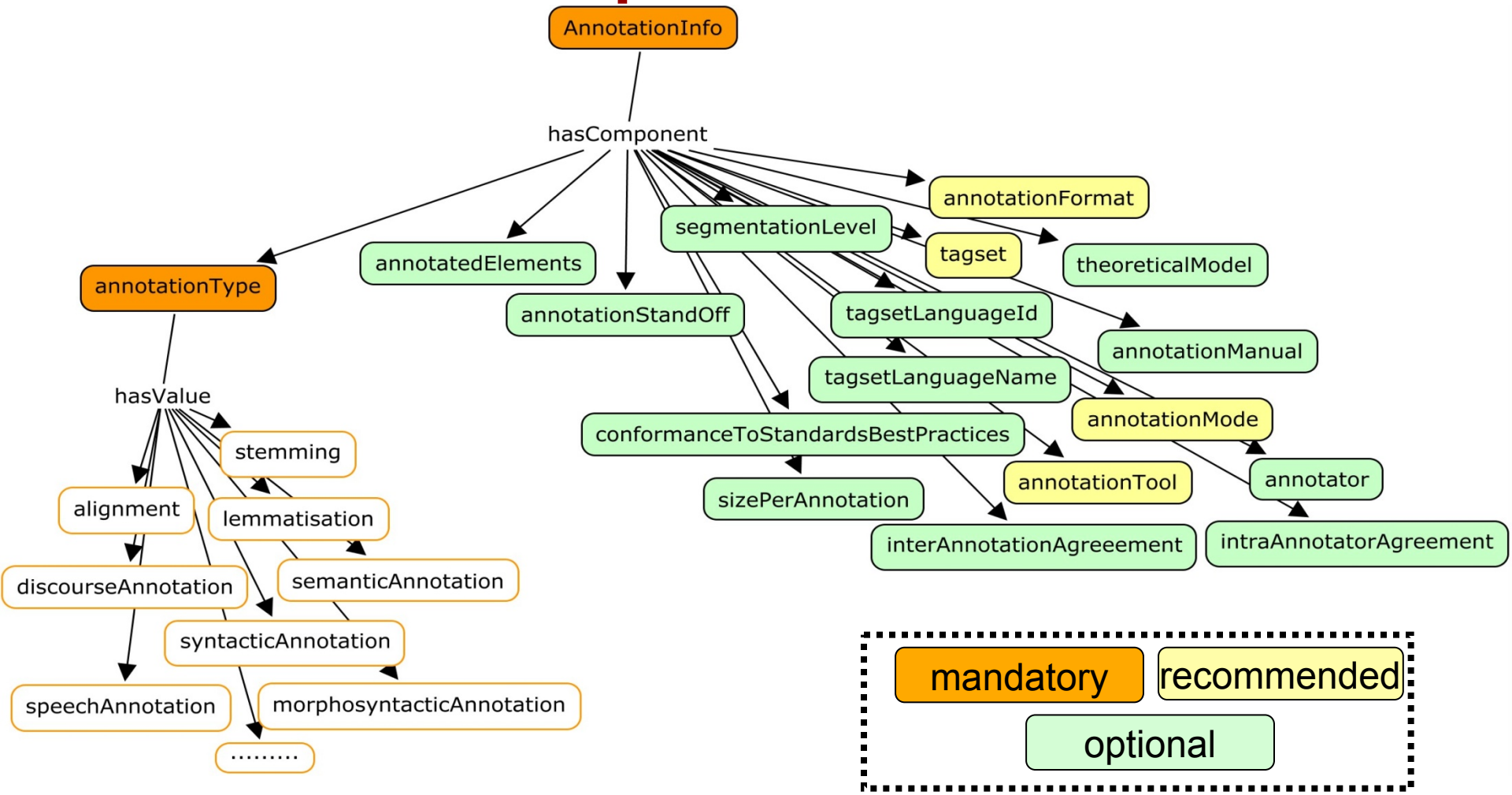
Identification component



corpusTextInfo



Annotation Component



Licensing issues

- Creative Commons
- META-SHARE Commons
  - CC-based licensing tool making LR available inside the network only.
  - for-free or for-a-fee
- META-SHARE “No Redistribution” licenses
  - permits the LR owner to have full control over the LR distribution.
  - for-free or for-a-fee
- Open Source Licenses proposed by META-SHARE: BSD, GPL, LGPL, Apache and AGPL.

Overview

- Background
- META-SHARE – basic features
- Applying META-SHARE – META-NORD user cases in Norway
- Observed challenges
- Conclusion

META-NORD contributions to META-SHARE

- 1st batch (Nov. 2011): http://metashare20.tilde.lv/
- 2nd batch (July 2012): will be using META-SHARE v2.1: http://metashare21.tilde.lv/

LRTs delivered for batch 1

Resource/tool	Resource Type	Availability	metadata provider
TRIS Spanish-German	Corpus	restrictedUse	META-NORD/UIB*
Parallel Treebank	Corpus	restrictedUse	META-NORD/UIB
Sofie Treebank	Corpus	restrictedUse	META-NORD/UIB
Scarrie Lexical Resource	Lexical resources	unrestrictedUse	META-NORD/UIB*
Norsk ordbank, Bokmål	Lexical resources	restrictedUse	Språkbanken
Norsk ordbank, Nynorsk	Lexical resources	restrictedUse	Språkbanken
Lexical database for Danish	Lexical resources	unrestrictedUse	Språkbanken
Lexical database for Norwegian	Lexical resources	unrestrictedUse	Språkbanken
Lexical database for Swedish	Lexical resources	unrestrictedUse	Språkbanken
Acoustic database for Danish	Speech resources	unrestrictedUse	Språkbanken
Acoustic database for Norwegian	Speech resources	unrestrictedUse	Språkbanken
Acoustic database for Swedish	Speech resources	unrestrictedUse	Språkbanken
Oslo-Bergen tagger	Tools	unrestrictedUse	META-NORD/UIB

Overview

- Background
- META-SHARE – basic features
- Applying META-SHARE – META-NORD user cases in Norway
- Observed challenges
- Conclusion

Language codes and searchability

- Language codes can be supersets of others.

- Consider the ISO-639-3 standard:
  - Norwegian: *nor*
  - Bokmål: *nob* (written norm)
  - Nynorsk: *nno* (written norm)

  A search for *nor* would not match records coded as *nob* or *nno*.

- Possible in META-SHARE (v. 2) to provide multiple language codes to monolingual resources.

SCARRIE
one LexicalResource, seven Lexicon elements with different properties

```
<LexicalResource dtdVersion="16">
<GlobalInformation>
 <feat att="languageCoding" val="ISO:639-3" />
 <feat att="license" val="CC-BY" />
 <feat att="authors" val="Victoria Rosén and Koenraad De Smedt (UiB), Torbjørn Nordgård (NTNU)" />
</GlobalInformation>

<Lexicon>
 <feat att="name" val="prefixes"/>
 <feat att="language" val="nob"/>

<LexicalEntry><Lemma />
<WordForm>
 <feat att="writtenForm" val="a" />
 <feat att="corrStyle" val="MN" />
 <feat att="morSynFeat" val="Pref,FREQUENT_AS_COMPOUND" />
</WordForm>
</LexicalEntry>

<LexicalEntry><Lemma />
<WordForm>
 <feat att="writtenForm" val="A" />
 <feat att="corrStyle" val="MN" />
 <feat att="morSynFeat" val="Pref,FREQUENT_AS_COMPOUND" />
</WordForm>
</LexicalEntry>
```

Complex resources

- We need a metadata scheme for complex resources, allowing for
  - separate metadata for every subpart that can be considered an individual resource.
  - grouping those metadata sets in the entry for the resource as a whole.
  - searching and retrieving all parts, or only the subpart that a user is looking for.

The juridical side

- **Rightholders**
  - Complex IPR relationships

    (original text owner, linguistic annotation owner, formal IPR holder,

    'moral right' holder)
- **Licensing**
  - Many resources need a restricted license even if the restrictions are minor (e.g. CC-BY).
- Precise guidelines and/or manual validation from an expert is pertinent.

Conclusion

- **Documentation:** Must be simple, intuitive and precise.
- **Consistency:** There should be a closer cooperation between developers of novel formats (e.g. LMF) and developers of catalogues and repositories.
- **Mandatory metadata description:** Lobby to make metadata descriptions of LRT mandatory in future funded projects.
- **A strategic dissemination plan**: Would ensure that resource providers and resource users know the META-SHARE metadata schema.

Links

- **META-SHARE documentation & user manual**

  http://www.meta-net.eu/meta-share/

- **META-SHARE knowledge base**

  http://metashare.ilsp.gr/portal/knowledgebase

- **META-SHARE user forum**

  http://www.meta-share.org/portal/forum/

Thank you for your attention!