



META-SHARE: An Open Resource Exchange Infrastructure

Elina Desipri
Athena RC, Greece
elina@ilsp.athena-innovation.gr

META **NORD**
IPR and Metadata Workshop



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).

META-SHARE: objectives

ETA-SHARE tries to treat problems and fill in gaps related to **visibility, documentation, identification, availability, preservation, interoperability**, etc. in the area of language data and (basic language processing) tools

t launches a rather **long-term endeavour** by which language resources are recognised as important assets that can **boost research, technology and innovation** through **wide availability, pooling, openness and sharing**

META-SHARE: what it is

ETA-SHARE is an **open, integrated, secure, and interoperable** language resources exchange infrastructure for the Human Language Technologies domain

marketplace where language data and tools are documented, uploaded and stored in repositories, catalogued and announced, downloaded, exchanged, discussed, aiming **to support a data economy** (free and for-a-fee LRs/LTs and services)

t brings together several organisations and initiatives



n language resources that are useful for driving applications useful to the EU citizen, primarily in his everyday communication and information search needs

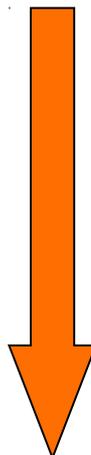
he vision of the multilingual information society

U languages and major world languages

Language Resources

- they serve the language technology development,
- they support the multilingualism and the linking between languages,
- they are being maintained/ supported to ensure extensibility,
- sustainability and repurposing,
- they are (ideally) sufficiently documented and described with standardized metadata schema,
- they are (ideally) available in the open domain

type of resources and technologies:

- 
- language data description, collection and cataloguing,
 - language processing tools description, collection and cataloguing,
 - evaluation packages
 - language data processing services through tools and technologies (starting from basic ones)
 - workflows by integrating language processing services

Resources and Media Types

annotated at any level,

text corpora,

morphological analyzers, multi-level (sentence-word-phrase) text aligners etc.),

analysers, semantic taggers, named entity recognizers etc.),

- monolingual **text** and **audio corpora**, raw and
- bi-/multilingual (comparable and parallel)
- mono-/ bi-/ multilingual **lexica**,
- basic language processing **tools** (tokenisers, various text analytics tools (syntactic
- **multimodal** corpora,
- audio and multimodal processing tools
- **language models** etc.



META-SHARE Metadata Model

Main features 1/2

descriptions of

- **LRs**, encompassing both **data** (*textual, multimodal/multimedia and lexical*) and **tools/technologies** used for their processing
- **related objects** (*reference documents, actors, activities etc.*)
- structure:
 - "**components**" are used to group together elements and relations (and other components)
 - elements linked to ISOcat Data Categories
 - "**profiles**" for each LR type are built upon components, incl.
 - components common to all LR types (e.g. identification, contact, distribution etc.)
 - LR type-specific components (e.g. annotation)

Main features 2/2

istinguishes:

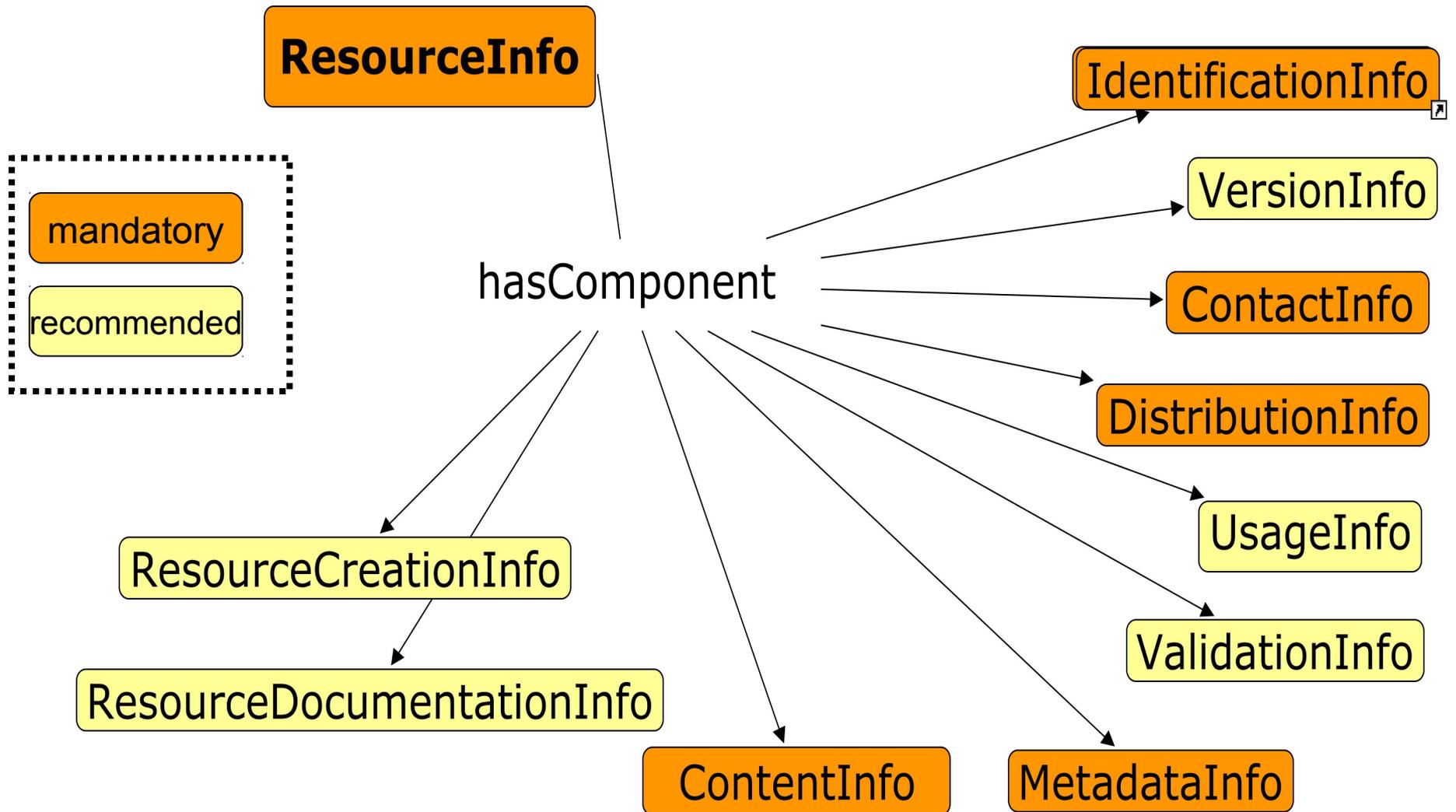
minimal schema:

- minimum set of obligatory elements and relations required for effective LR search, identification and retrieval
- e.g.: identification info (title, unique identifier), contact details, technical info (language, size etc.), distribution info

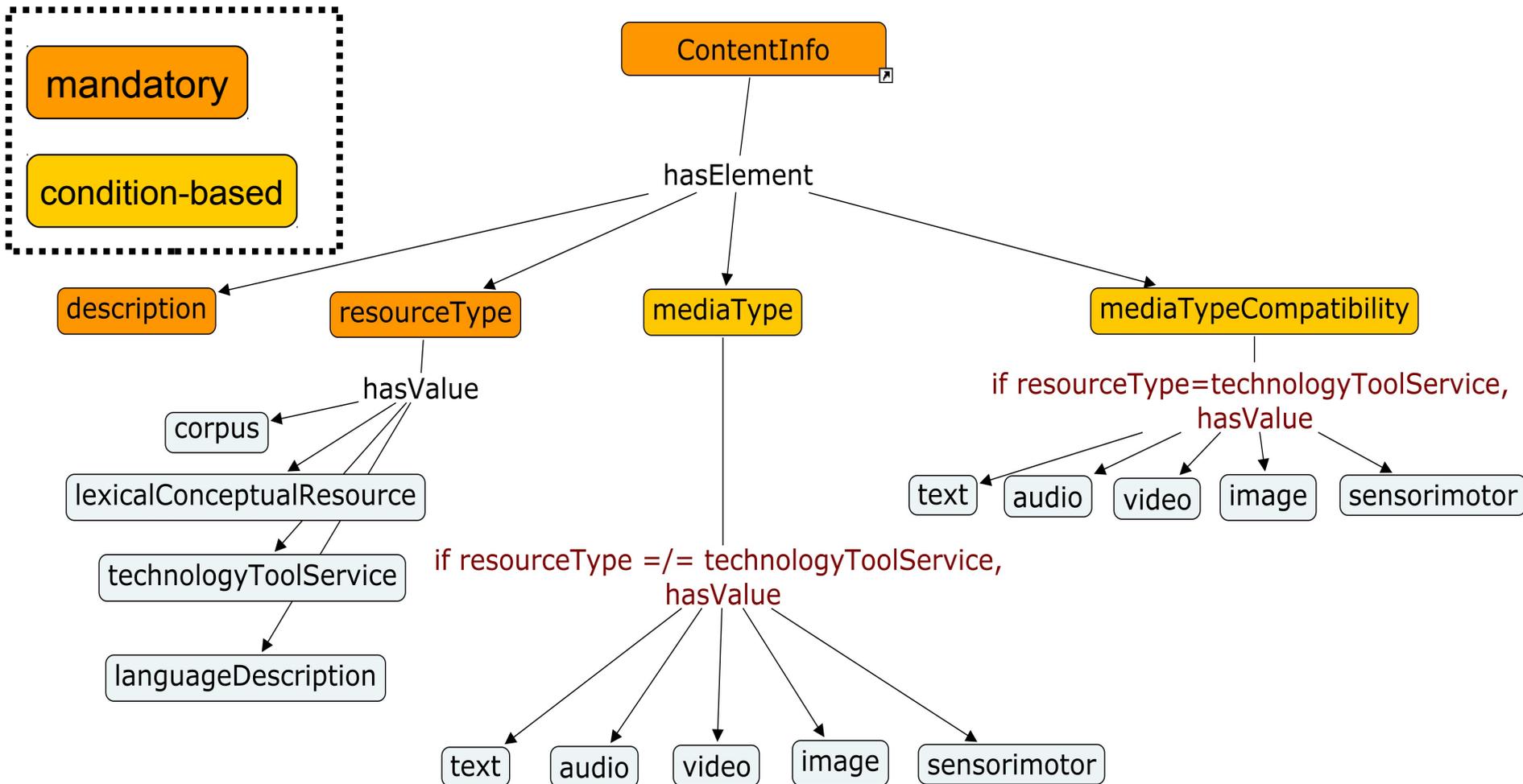
▪ ***maximal schema***

- all elements and relations required for the description of LRs (i.e. additional set of recommended and optional elements and relations for the whole lifecycle of LR production and usage)
- e.g.: provenance information, creation details, validation/evaluation information, intended and actual use(s) etc.

Resource Info

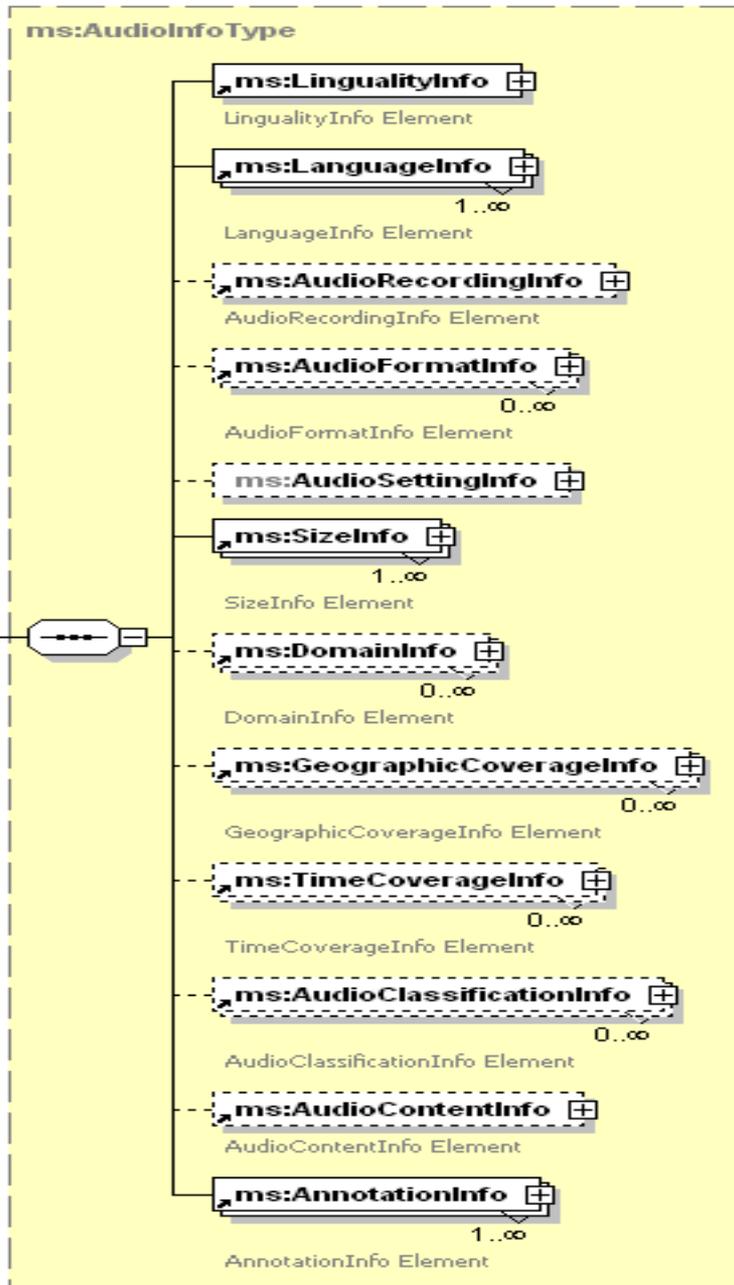


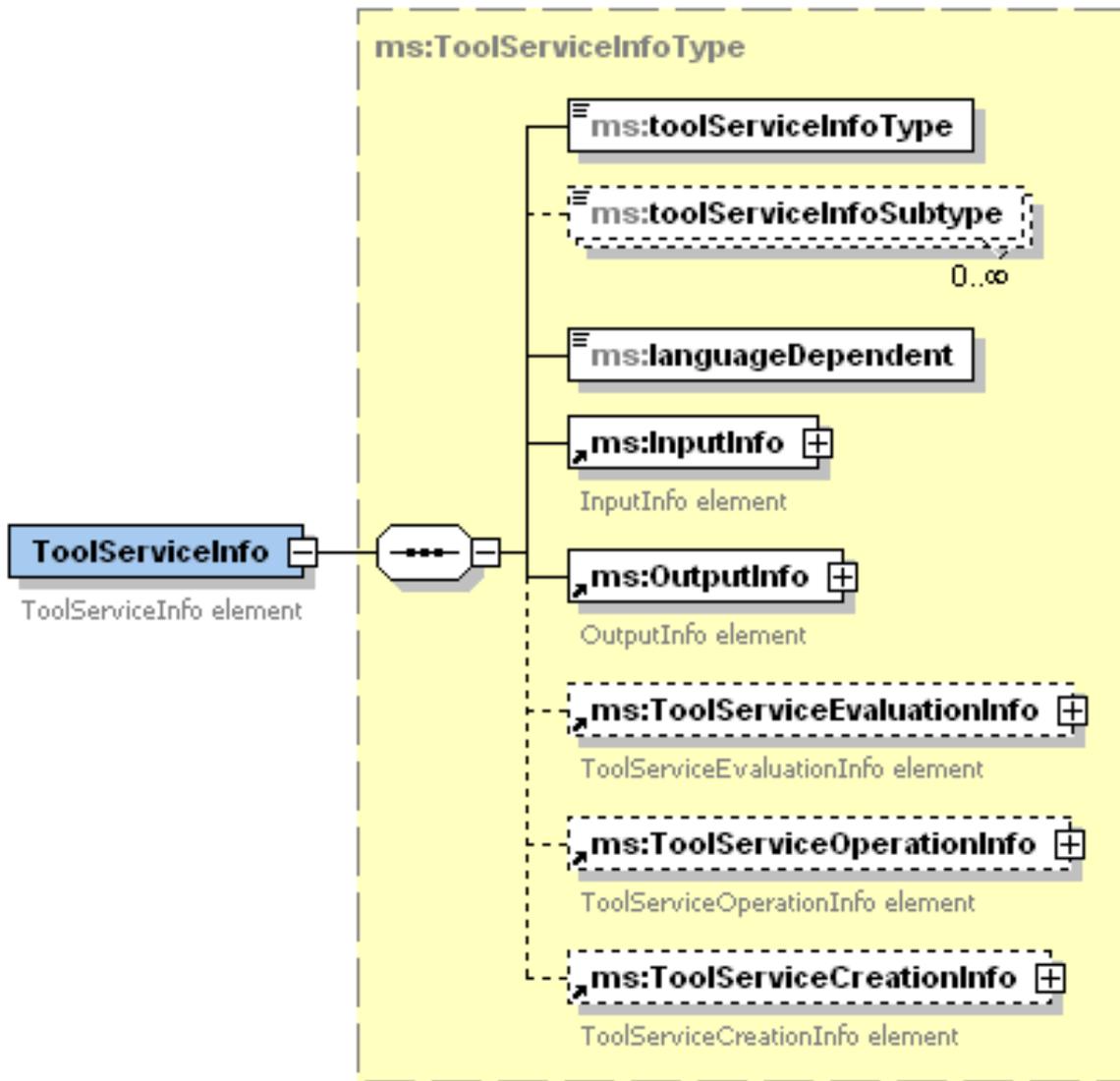
Content component

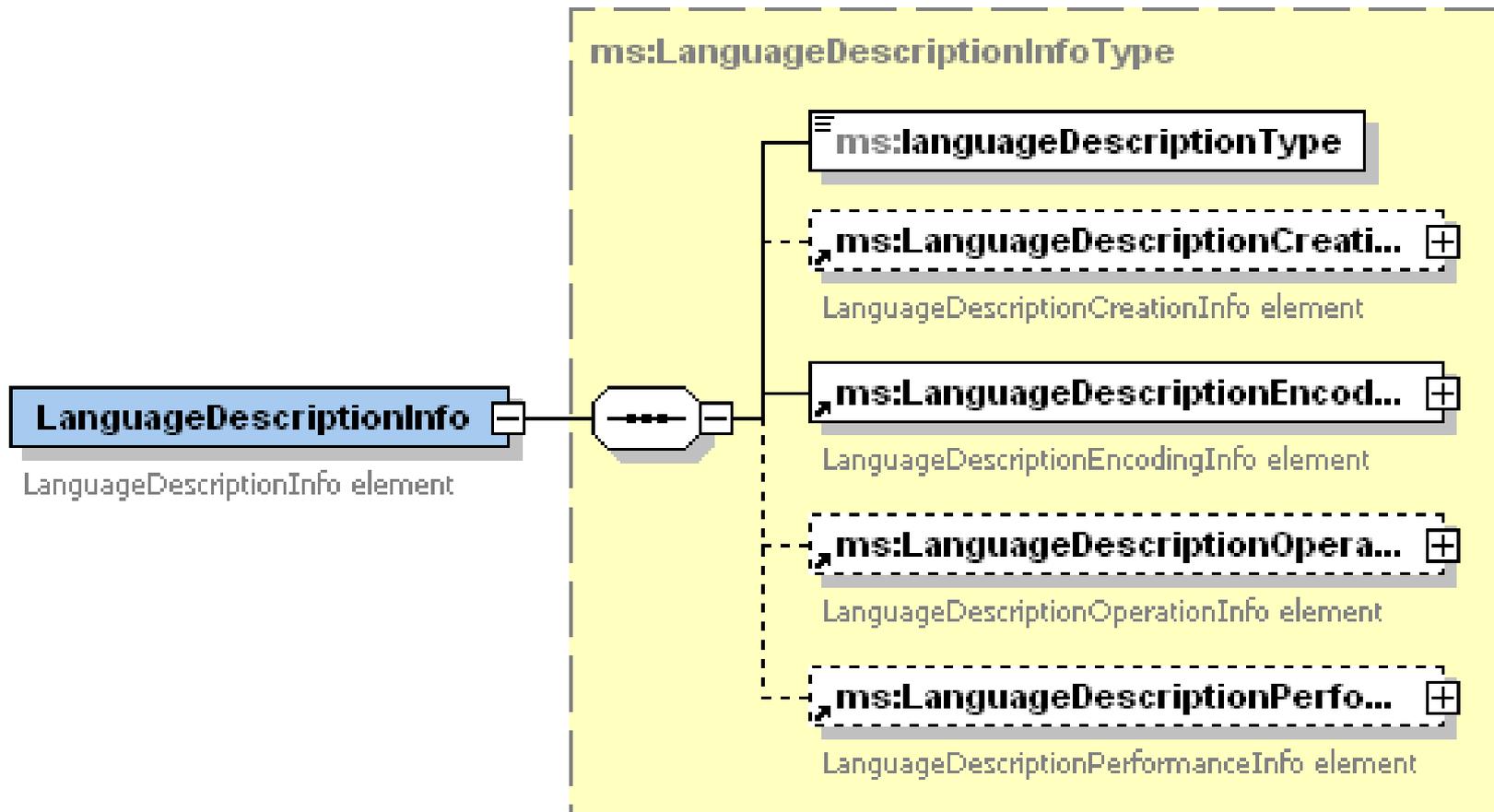


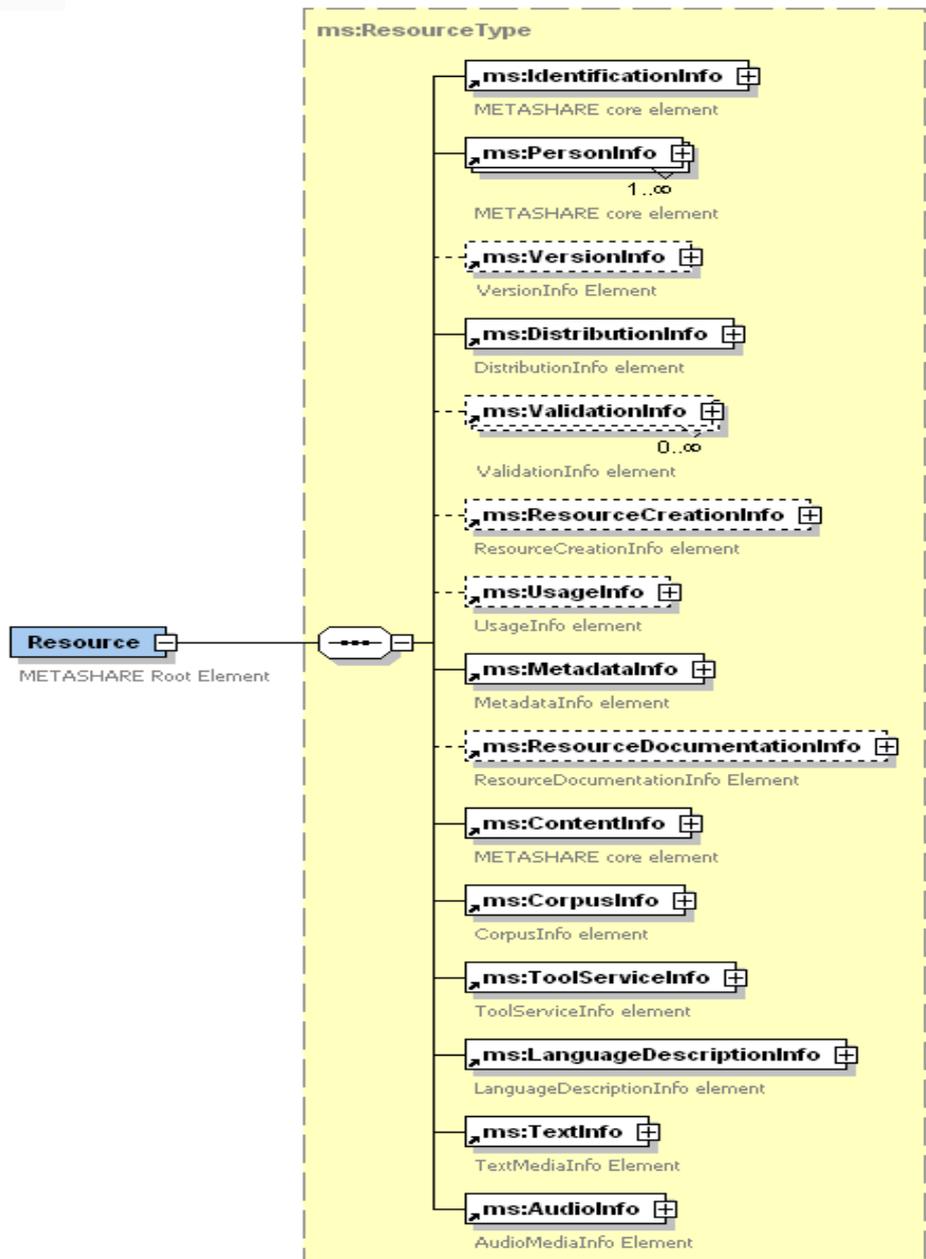
Progress report

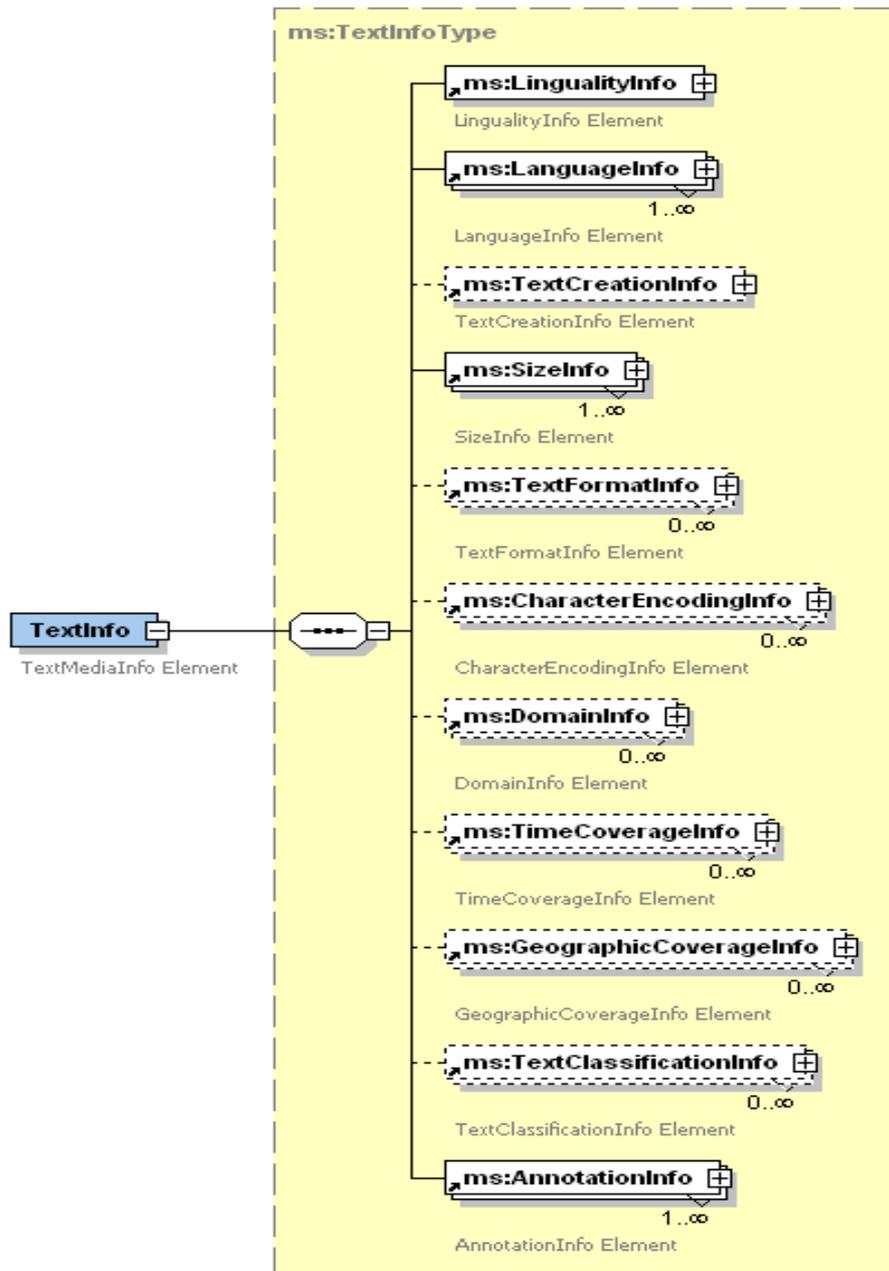
Language Resource Type and Medium	MD model	META-SHARE implementation
Text Corpora	✓	✓
Lexical Conceptual Resources	✓	✓
Audio Corpora	✓	in progress
Video corpora	✓	in progress
Tools/Services	✓	in progress
Language Descriptions	✓	in progress
Image	in progress	in progress
Sensorimotor	in progress	in progress











META-SHARE version 1

www.meta-share.eu,

www.meta-share.org

www.meta-net.eu/meta-share

ETA-SHARE Version 1: September 2011

- working version of META-SHARE to be rolled out within the META-NET network
- focus on laying the distributed repository infrastructure by making available free sw for repository setup to help organise, document LRs
- basic principles and concepts of the MD model, and its implementation as regards written corpora and lexical resources, for validation and exemplification purposes

ETA-SHARE Version 2

- stable version, ready for production use
- focus on providing services to LR providers, consumers and aggregators (statistics, reporting, recommendations)



HANK YOU!