META-SHARE and LOD

Stelios Piperidis

Institute for Language and Speech Processing/ILSP
Athena Research Center

LIDER Workshop
Madrid, 9 May 2014

META-SHARE is …

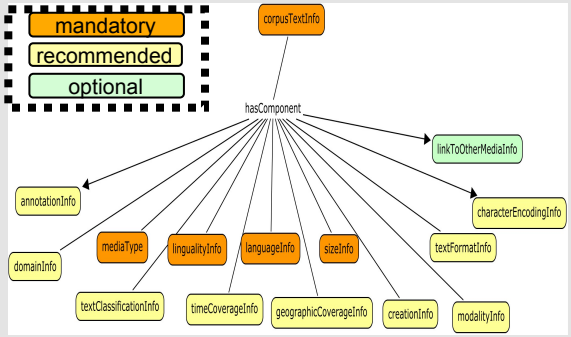
- an open infrastructure, a network of distributed repositories for language resources (datasets and language processing tools)
- the counts
  - a network of 34 organisations
  - 27 repositories
  - > 2.400 datasets/tools/services (master copies)
  - > 2.600 downloads of resources
  - > 14.000 user sessions
  - > 4.500 update actions

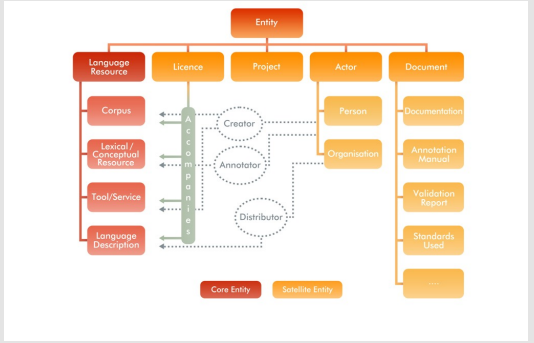
META-SHARE & QT21

- For the needs of the QTLaunchPad project, a new MT dedicated META-SHARE compliant repository : QT21 with a new layer of language processing functionalities, provided as web services
  - from simple services to pipelines of services (workflows)
- enabling
  - linking of datasets with language processing services
  - processing/annotation of datasets by relevant services
  - provenance tracking (through linking raw and annotated version(s) of datasets and associated info)

META-SHARE metadata model

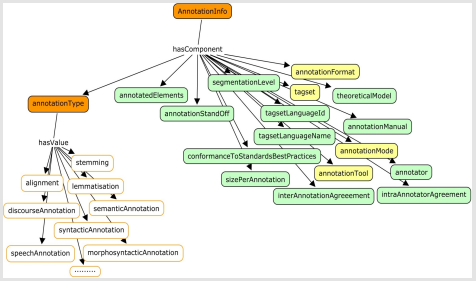
- Long tradition of metadata standardisation attempts
- We did not start from scratch:
  - based on overview of other schemas
  - result of consultation with a large number of experts
  - catering for all resource (corpora, lexica, language descriptions, tools/ services) and media types (text, audio, image, video), and subtypes (n-gram corpora, sensorimotor data and related measurements)
  - formalised XSD schema
  - organised in components linking semantically coherent elements
- Glue between repositories through harvesting
- OAI-PMH bridge for linking with other infrastructures





META-SHARE metadata model issues-challenges-improvements

- Persistent identification
  - Unit of identification (e.g. file-collection-corpus, entry-lexicon)
  - Versioning (datasets and tools)
  - Replicability
- Relations and their representation
  - Dataset - dataset
  - Dataset - tool/service/workflow
  - Tool/service/workflow - tool/service/workflow



- Linking to external resources – controlled vocabularies, other data repositories, LOD, …
- Complexity – e.g. sheer number of licences with non-compatible specification of permissions/obligations/…
- Enrichment of tool/service & workflow descriptions (cf. QT21)
- Inconsistencies and bug fixing

Links

- META-SHARE metadata schema rdf-ification experiment (UPF)
  - simpler RDF model
  - integrating Panacea/BioCatalogue registry
  - facilitating internal and external linking
  - initial experiment on UPF datasets
  - converter now being tested on the whole inventory with an OAI-PMH dump through the ILSP managing node
  - validation of conceptual principles
  - for the near future, need to maintain both models (XML/RDF)

- for the purposes of full-blown language infrastructures, specifically for interoperability & annotation merging and exchange, keep an eye on NIF and NLP2RDF developments