Standards for language encoding: Sharing resources

Tomaž Erjavec
Dept. of Knowledge Technologies
Jožef Stefan Institute

ESSLLI 2011

Sharing language resources

- Copyright
- Making information about resources available: metadata + harvesting
- Making lingustic categories used by the resources harmonised: ontologies

Overview of the lecture

1. Copyright
2. Meta-data: TEI, DC, OAI-PHM
3. Semantic Web:
   RDF, RDFs OWL (W3C standards)
4. EU projects

Copyright, copyleft

- Text is by definition copyrighted; big problem for compiling and sharing e.g. corpora
- Open Source: make programs available in source code
- With corpora: make them available as „source code“; Open data
- Creative Commons licences
- e.g. JOS corpora

Metadata

- Data about data: the description of a language resource
- Problem: arrive at a commonly accepted set of metadata descriptors and the means to exchange them

TEI header

- **<fileDesc>** *file description*,
  Bibliographical description of the file itself & about the source(s) from which the electronic text was derived.
- **<encodingDesc>** *encoding description*,
  Relationship between the electronic text and its source(s): text normalisation, levels of encoding, analysis, etc.
- **<profileDesc>** *text profile*
  Classificatory and contextual information about the text, e.g. its subject matter. In corpora it is desirable to enforce a controlled descriptive vocabulary (text type and origin).
- **<revisionDesc>** *revision history*,
  History of changes made during the development of the electronic text. For version control and resolving questions about the history of a file.

Example teiHeader

Some metadata standards

- Books are the oldest things to have metadata standards:
  - MARC - MAchine-Readable Cataloging (also MARC XML)
    (for traditional, printed books and publications)
  - MODS: Metadata Object Description Schema
    (less complex than MARC)
  - METS: Metadata Encoding and Transmission Standard
    (for digital libraries)
- General MD standards:
  - Dublin Core
    (core metadata descriptors)
    - OAI-PMH
      (Open Archives Initiative Protocol for Metadata Harvesting)

Persistent Identifiers

- URN: ISO standard Uniform Resource Name
- PURL: Persistent URLs
  a redirection service, e.g.
  http://purl.org/olia/mte/multext-east.owl →
  http://nl.ijs.si/ME/owl/multext-east.owl
- DOI: ISO standard Digital Object Identifier
  a character string to identify an electronic document
  - scholarly materials (journal articles, books, etc.)
  - scientific data sets
  - EU official publications
- Attempts to standardise identifiers for LRs:
  ISO TC 37: ISO/WD 24618 -- Citation of Electronic Resources

Dublin Core Metadata

- The Dublin Core Metadata Initiative, or "DCMI", is an open organization engaged in the development of interoperable metadata standards
- The story starts with OCLC/NCSA Metadata Workshop, 1995, Dublin, Ohio USA
- Proposal for "core metadata" for simple and generic resource descriptions: "Dublin Core", 15 elements
- Now widely adopted
  - IETF RFC 5013
  - ANSI/NISO Standard Z39.85-2007
  - ISO Standard 15836:2009

Dublin Core Metadata Element Set (DCMES)

DMCES defines 15 elements:

- *Title*: A name given to the resource.
- *Creator*: An entity primarily responsible for making the content of the resource.
- *Subject*: The topic of the content of the resource.
- *Description*: An account of the content of the resource.
- *Publisher*: An entity responsible for making the resource available.
- *Contributor*: An entity responsible for making contributions to the content of resource.
- *Date*: A date associated with an event in the life cycle of the resource.
- *Type*: The nature or genre of the content of the resource.
- *Format*: The physical or digital manifestation of the resource.
- *Identifier*: An unambiguous reference to the resource within a given context.
- *Source*: A Reference to a resource from which the present resource is derived.
- *Language*: A language of the intellectual content of the resource.
- *Relation*: A reference to a related resource.
- *Coverage*: The extent or scope of the content of the resource.
- *Rights*: Information about rights held in and over the resource.

DCMES in XML

```
<?xml version="1.0"?>
<metadata xmlns="http://example.org/myapp/"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <dc:title>UKOLN</dc:title>
  <dc:description>UKOLN is a national focus ...</dc:description>
  <dc:publisher>UKOLN, University of Bath</dc:publisher>
  <dc:identifier>http://www.ukoln.ac.uk/</dc:identifier>
</metadata>
```

Further developments

- more MD elements added, e.g. abstract, accessRights, accrualMethod, …
- Each element (term) is described using Semantic Web concepts, e.g.

Term Name: accrualMethod	
URI:	http://purl.org/dc/terms/accrualMethod
Label:	Accrual Method
Definition:	The method by which items are added to a collection.
Type of Term:	Property
Has Domain:	http://purl.org/dc/dcmitype/Collection
Has Range:	http://purl.org/dc/terms/MethodOfAccrual
Version:	http://dublincore.org/usage/terms/history/#accrualMethod-003

OAI-PMH

- **OAI**: Open Archives Initiative
  Enabling access to Web-accessible material through interoperable repositories for metadata sharing, publishing and archiving
- **OAI-PMH**: OAI Protocol for Metadata Harvesting
  a lightweight harvesting protocol for sharing metadata between services: HTTP + XML + DC
- **Harvesting**
  gathering metadata from distributed repositories into a data store
- **Data Provider**
  maintains repositories (web servers) that support the OAI-PMH as a means of exposing metadata.
- **Service Provider**
  issues OAI-PMH requests to data providers and uses the metadata as a basis for building value-added services.

OLAC

- OLAC: Open Language Archives Community
- partnership of institutions and individuals creating a worldwide virtual library of language resources by:
  - developing consensus on best current practice for the digital archiving of language resources
  - developing a network of interoperating repositories and services for housing and accessing such resources
- uses OAI-PMH
- supports search over harvested metadata via OLAC Search Service
- Example data provider: http://nl.ijs.si/ (although not yet registered to OLAC)

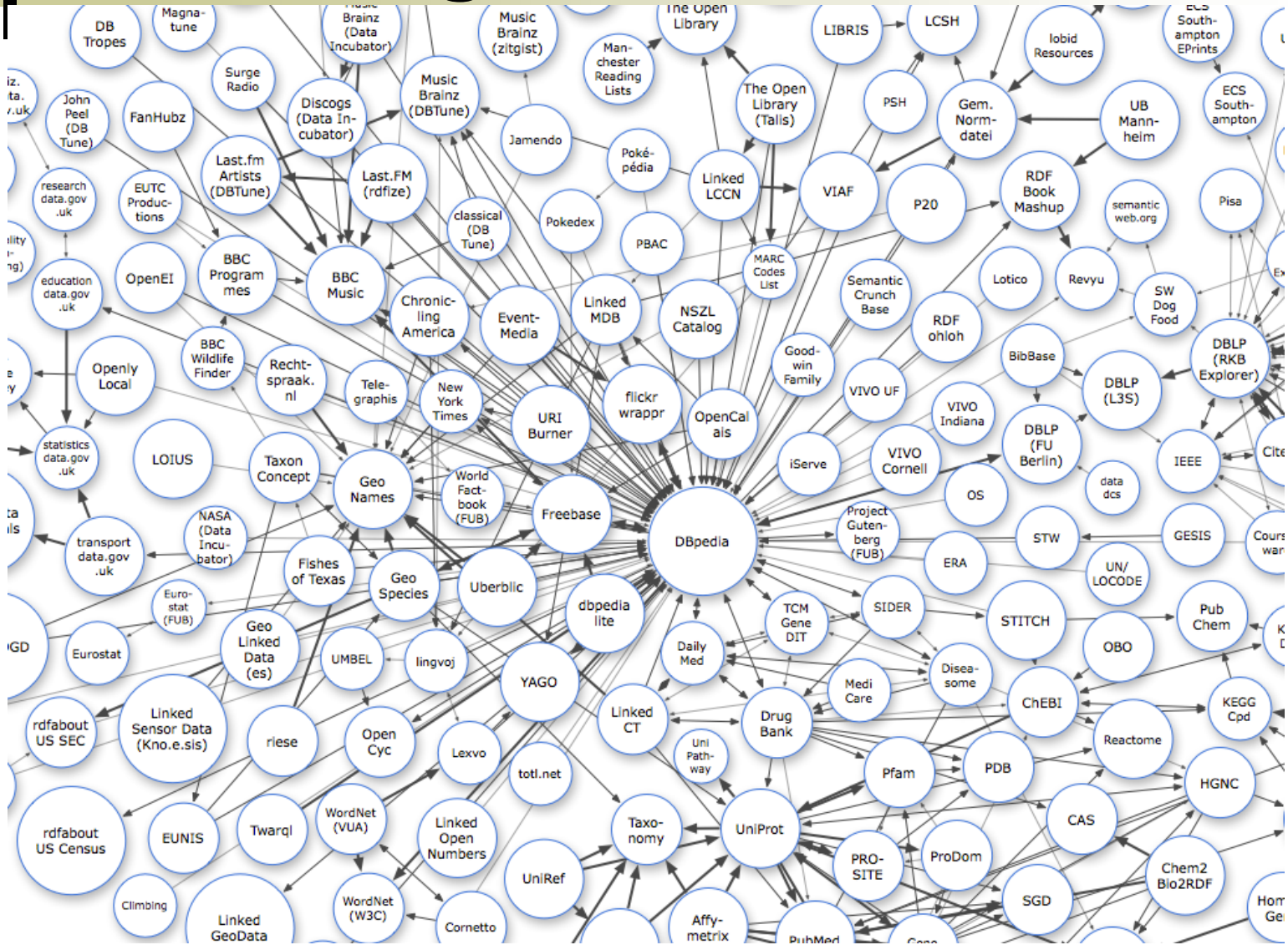
II. Semantic Web

- Idea by Tim Berners-Lee:
  Facilitate machines to understand the semantics of information on the WWW. It extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other, enabling automated agents to access the Web more intelligently and perform tasks on behalf of users.
- Very influential, many projects, W3C standards, technologies
- Semantic Web Technologies: RDF, RDFS, OWL
- Concept of „Linked data“

Linked Data

- Use URIs as names for things
- Use HTTP URIs (URLs) so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
- Include links to other URIs so that they can discover more things.

The Linking Open Data cloud diagram



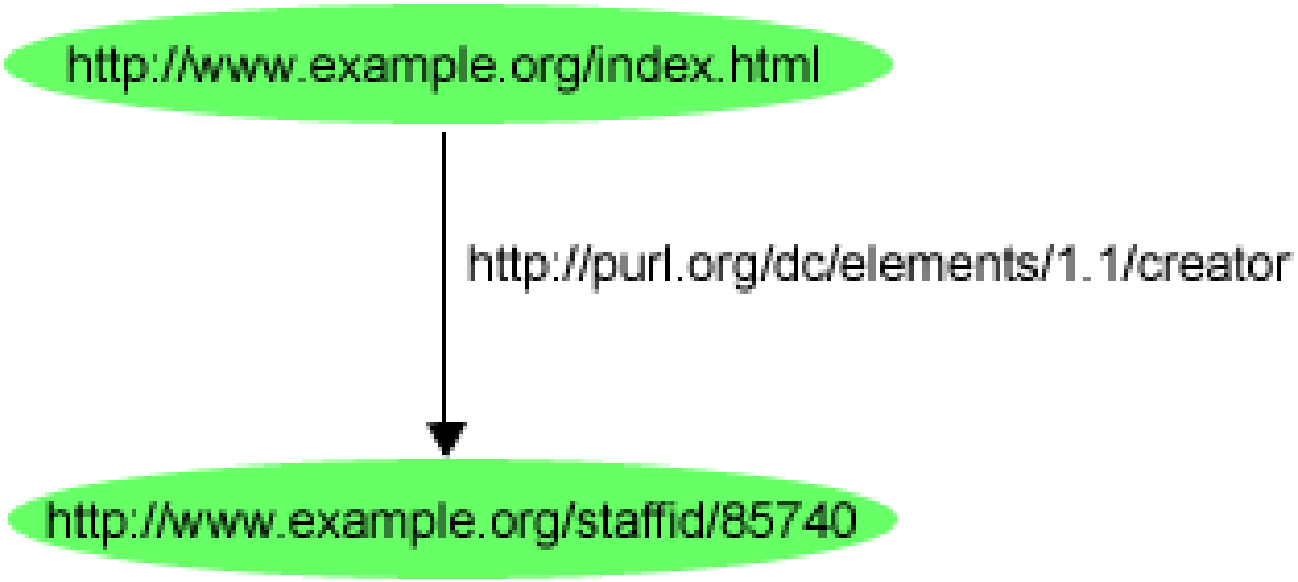
RDF

- Resource Description Framework (RDF) is a family of W3C specifications designed as general (meta)data model
- Similar to Entity-Relationship or Class diagrams: based on the idea of making statements about resources (in particular Web resources) in the form of **subject-predicate-object** expressions
- These statements are known as triples
- Can be expressed in many different formats

RDF triples

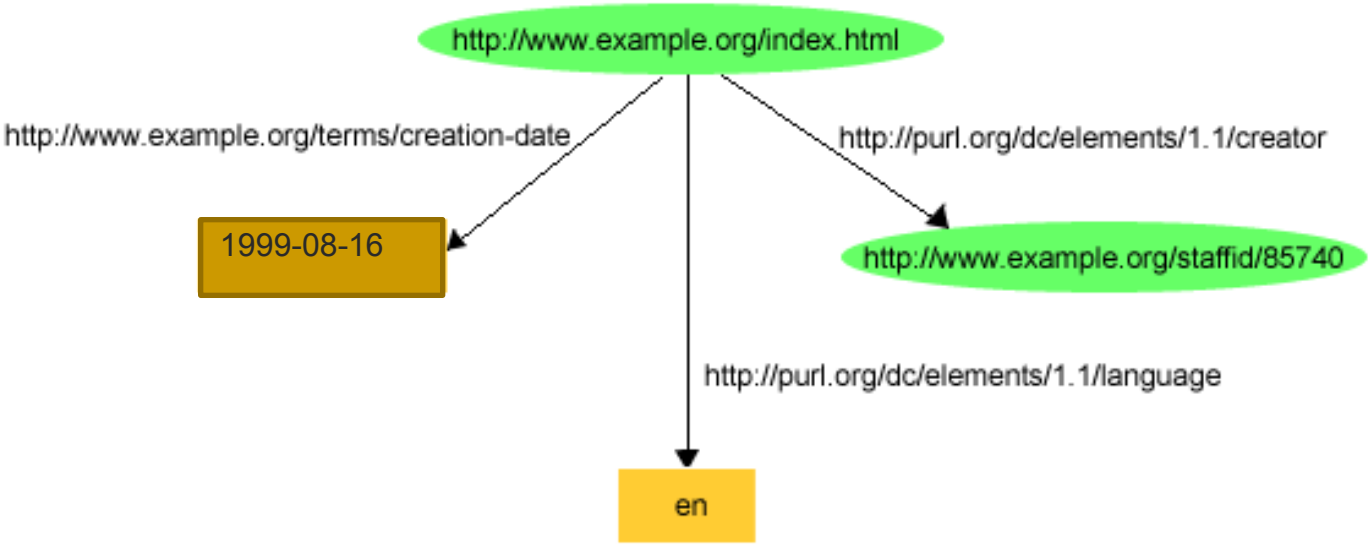
- Example of a statement about a Web page: *http://www.example.org/index.html* has a *creator* whose value is *John Smith*
- RDF terms for the 3 parts of these statement are:
  - **subject**: *http://www.example.org/index.html*
  - **predicate**: *creator*
  - **object**: *John Smith*
- All 3 are **URI references**, e.g.
  - subject URI, e.g. *http://www.example.org/index.html*
  - predicate URI, e.g. *http://purl.org/dc/elements/1.1/creator*
  - object URI, e.g. *http://www.example.org/staffid/85740*

As a graph:


http://www.example.org/index.html
http://purl.org/dc/elements/1.1/creator
http://www.example.org/staffid/85740

Literals

- More statements about the Web page:
  - *http://www.example.org/index.html* has a *creator* whose value is *John Smith*
  - *http://www.example.org/index.html* has a *creation-date* whose value is *August 16, 1999*
  - *http://www.example.org/index.html* has a *language* whose value is *English*
- Objects (but not subjects or predicates) in RDF statements can also be constant values, called *literals*



Namespaces

- Namespaces (qualified names) can be used to make statements shorter
- Given namespace definitions:
  - dc: http://purl.org/dc/elements/1.1/
  - ex: http://www.example.org/
- We can write
  - ex:index.html dc:creator ex:staffid/85740 .
    ex:index.html ex:terms/creation-date "1999-08-16" .
    ex:index.html dc:language "en" .

RDF in XML

- **RDF triplet:**
  ex:index.html ext:creation-date "1999-08-16" .

- **XML syntax:**

```
<?xml version="1.0"?>
 <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
                xmlns:ext="http://www.example.org/terms/">
   <rdf:Description rdf:about="http://www.example.org/index.html">
     <ext:creation-date>1999-08-16</ext:creation-date>
   </rdf:Description>
</rdf:RDF>
```

Literals

- Literals need not be only strings:
  - they can have assigned types (e.g. integer, date, URI)
  - they can be complex:
    - bag, set, alt
    - structured values (think feature-structures)
    - fragments of XML

RDF Summary

- To describe resources RDF uses triples subject/predicate/object (resource/property/value)
- Their values are URI references, and, for objects, literals
- All other features (types, collections,...) are just syntactic sugar
- RDF is simply a description language - it does not specify how to interpret these descriptions
- The basic RDF model is in terms of graphs - RDF/XML is only a representation of such graphs

RDFS

- RDF users need the ability to define the vocabularies (terms) they intend to use in RDF statements, to indicate that they are describing specific kinds or classes of resources, and will use specific properties in describing those resources
- RDF itself provides no means for defining such application-specific classes and properties
- Such classes and properties are described as an RDF vocabulary, using extensions to RDF provided by the *RDF Vocabulary Description Language: RDF Schema*

RDFS Classes and subclasses

- Classes identify the various kinds of things to be described
- A (sub)class is any resource having an rdf:type property whose value is the resource rdfs:Class/rdfs:subClassOf, e.g.

```
ex:MotorVehicle          rdf:type          rdfs:Class .
ex:PassengerVehicle      rdfs:subClassOf   ex:MotorVehicle .
exthings:companyCar      rdf:type          ex:PassengerVehicle
```

- Example: http://purl.org/dc/terms/accrualMethod

Ontologies

- AI definition: An ontology is *a formal specification of a conceptualisation*
- The language used to describe the knowledge domain should not only have a formal syntax, but also a formal semantics
- The fact that a formal semantics is defined enables *reasoning* over an ontology, i.e. new, valid statements can be deduced on the basis of explicit statements of the ontology

OWL

- OWL is a language for defining and instantiating Web ontologies and is a W3C Recommendation
- OWL is built upon XML, XML Schema, RDF and RDF Schema
- OWL differs from RDFS in that it is a knowledge representation, not a message format, i.e. it also supports reasoning over data
- The formal foundation of OWL is provided by Description Logics
- OWL was designed to support distributed knowledge sources

Three Species of OWL

- *OWL Lite* supports classification hierarchies and simple constraints and provides a quick migration path for taxonomies.
- *OWL DL* gives maximum expressiveness while retaining computational completeness (all conclusions are guaranteed to be computable) and decidability (all computations will finish in finite time).
- *OWL Full* gives maximum expressiveness and the syntactic freedom of RDF with no computational guarantees.
- Each of these sublanguages is an extension of its simpler predecessor, both in what can be legally expressed and in what can be validly concluded.

Open World Assumption

- In DBs, we typically make a closed-world assumption:
  If a person is not listed in the employee DB, that person is not an employee
- But the Semantic Web is inherently distributed, so any information is potentially incomplete - what is not asserted might nevertheless be true
- OWL makes an open world assumption, i.e. descriptions of resources are not confined to a single file or scope. While class C1 may be defined originally in ontology O1, it can be extended in other ontologies.
- The consequences of additional propositions about C1 are monotonic. New information cannot retract previous information. New information can be contradictory, but facts and entailments can only be added, never deleted.
- The possibility of such contradictions is something the designer of an ontology needs to take into consideration.

Basic OWL Elements

- Most of the elements of an OWL ontology concern **classes, properties, instances of classes, and relationships between these instances**.
- Many uses of an ontology will depend on the ability to reason about individuals. So we need to have a mechanism to describe the classes that individuals belong to and the properties that they inherit by virtue of class membership. We can always assert specific properties about individuals, but much of the power of ontologies comes from class-based reasoning.
- The world of classes and individuals would be pretty uninteresting if we could only define taxonomies. Properties let us assert general facts about the members of classes and specific facts about individuals.

Example ontology elements

```
<owl:Class rdf:ID="ConsumableThing"/>

<owl:Class rdf:ID="Wine">
   <rdfs:subClassOf rdf:resource="#ConsumableThing" />
      ...
</owl:Class>

<owl:ObjectProperty rdf:ID="madeFromGrape">
  <rdfs:domain rdf:resource="#Wine"/>
  <rdfs:range rdf:resource="#WineGrape"/>
</owl:ObjectProperty>

<Wine rdf:ID="MikesFavoriteWine">
   <owl:sameAs rdf:resource="#StGenevieveTexasWhite" />
</Wine>
```

Ontologies and LRs

So what does all this have to do with LRs?

- Ontologies enable reasoning over concepts, and concepts are expressed in language
  - Monnet EU project: ontology lexicalisation & localisation
- If data categories of LRs are put in an ontology, we can reason over these properties; we can also link categories from different LRs together

Examples

- Already saw GOLD ontology
- Lexvo ontology (Gerard de Melo, MPI)
  information about language-related entities for the Linked Data Web and Semantic Web. The information is not only highly interconnected but also linked to a variety of resources on the Web.
- OLiA ontologies (Christian Chiarcos, Potdsdam)
  (integration of linguistic terminologies, mainly morphosyntax)
  - MULTEXT-East morphosyntactic specifications in OWL: http://nl.ijs.si/ME/owl/

EU LR standardisation projects

- Over the years many EU projects had as their goal standardisation of language resources
- First broad overview: EAGLES (Expert advisory Group on Language Engineering Standards)
- continued with ISLE, ...,
- ILC CNR in Pisa as the coordinator of many of these projects

Current EU LR projects

- A vision for future HLT (LR) research and goals:
  - FLaReNet: Fostering Language Resources Network
- Language Resource Sharing
  - META-NET: creating META-SHARE, “a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata”
  - Clarin: LRs for the Humanities
- Language Resource Construction
  - ACCURAT, PANACEA, TTC, …

FlaReNet

- Fostering Language Resources Network (2008-2011)
  „A major condition for the take-off of the field of Language Resources and Language Technologies is the creation of a shared policy for the next years.
  FLaReNet aims at developing a common vision of the area and fostering a European strategy for consolidating the sector, thus enhancing competitiveness at EU level and worldwide.“
- Yearly conferences; 2011, Venice
- FLaReNet Blueprint of Actions and Infrastructures

META-SHARE

- Part of the META-NET network
- META-SHARE is building a multi-layer infrastructure that will:
  - make available quality documented LRs and related metadata over the net,
  - ensure that such LRs and metadata are managed, preserved and maintained,
  - provide a set of services to all META-SHARE members and users,
  - promote the use of standards for LR building for maximum interoperability,
  - allow third parties to export their LRs over the META-SHARE network,
  - allow potential users of the LRs to easily and legally safely acquire the LRs requested for their own purposes.

To conclude…

- The nice thing about standards is that there is so many of them!
- Basic standards that you should understand and use:
  - Unicode, XML, language codes, dates and times
  - Depending on what you do:
    - TEI
    - DC
    - Semantic Web standards: RDF* (graph based methods in linguistics)
- A word of caution: not every standard is perfect, and not all standards get widely adopted
  - so, wait until they do, and they get software support

Thanks

for sticking with it!