
# Metadata

# Hands-on session

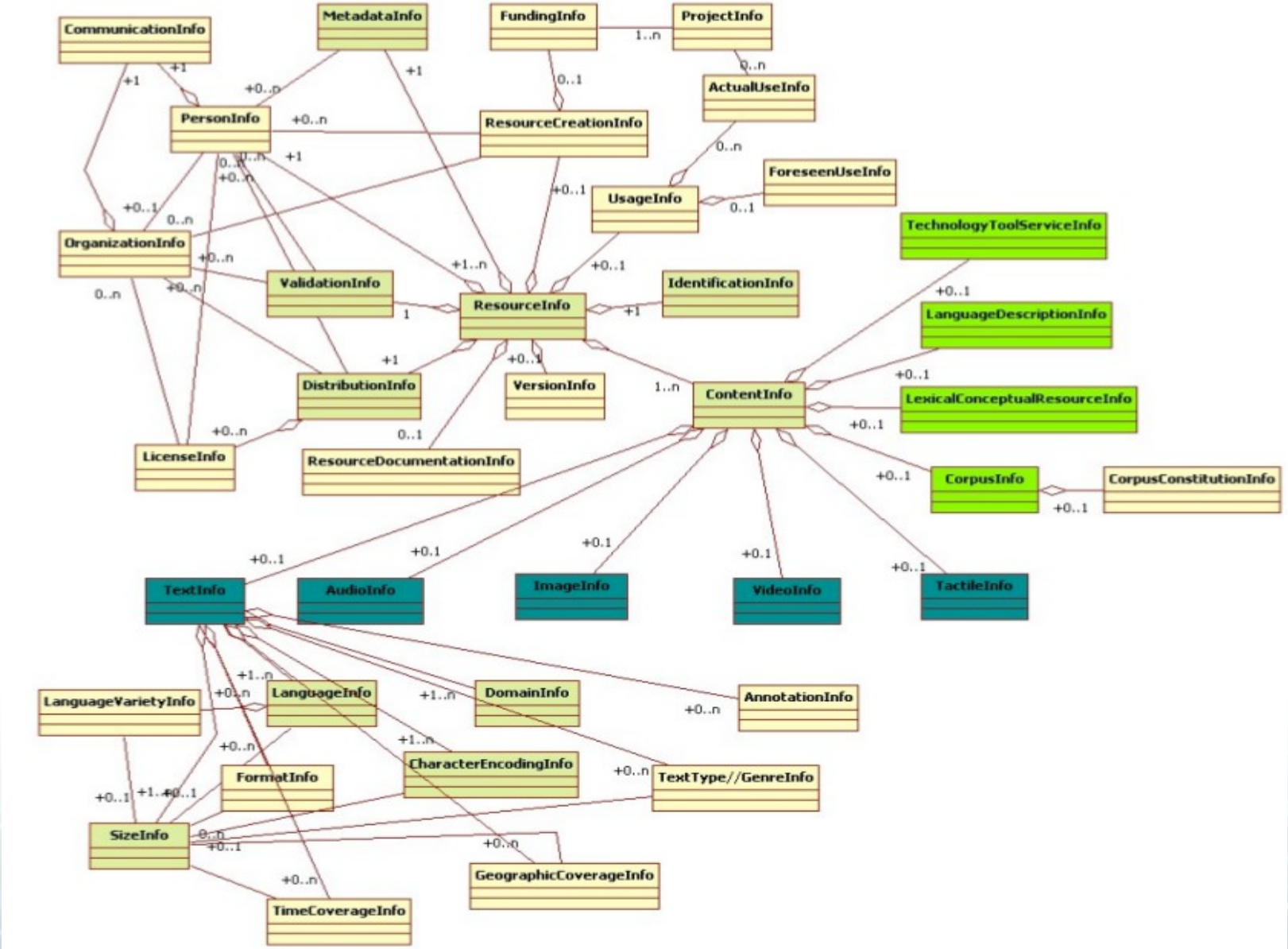
Markus Forsberg, Språkbanken, UGOT
Workshop in Helsinki, 2011-09-30

# Metadata (hands-on)

- Our job is to produce metadata and to upload the first batch by the end of November 2011.
- The metadata format is still in flux, and it is not clear what uploading actually means.
- However, we still need to do our job and produce (reasonable) metadata.
- Getting there:
  - Quick introduction+example (*Specification of Metadata-Based Descriptions for Language Resources and Technologies, March 17, 2011*)
  - Hands-on session
  - Metadata discussion in the afternoon with a focus on points to raise at the META-SHARE metadata workshop in Athens

# Minimal/maximal schema

- Minimal schema – what is required
  - Mandatory (M)
  - Condition-dependent Mandatory (MC)
- Maximal schema – what is possible to express
  - Recommended (R)
  - Optional (O)
- We will focus on the minimal schema


CommunicationInfo
MetadataInfo
FundingInfo
ProjectInfo
ActualUseInfo
PersonInfo
ResourceCreationInfo
ForeseenUseInfo
UsageInfo
OrganizationInfo
TechnologyToolServiceInfo
ValidationInfo
IdentificationInfo
ResourceInfo
LanguageDescriptionInfo
DistributionInfo
VersionInfo
ContentInfo
LexicalConceptualResourceInfo
LicenseInfo
ResourceDocumentationInfo
CorpusInfo
CorpusConstitutionInfo
TextInfo
AudioInfo
ImageInfo
VideoInfo
TactileInfo
LanguageVarietyInfo
LanguageInfo
DomainInfo
AnnotationInfo
CharacterEncodingInfo
FormatInfo
TextType//GenreInfo
SizeInfo
GeographicCoverageInfo
TimeCoverageInfo

# Minimal schema+Example

- A run-through of the minimal schema and how we chose to describe two of the resources at Språkbanken.
- The two resources:
  - SALDO – a lexical-semantic resource
  - BLOGGMIX – a dynamic corpus
- (Red font = not part of D2.2)

# Minimal schema – overview 1

- *IdentificationInfo*

  resourceTitle

  pid

  identifier

- *DistributionInfo*

  availability

  licenseInfo

  distributionMedium

- *ContentInfo*

  description

  resourceType

  mediaType *(text, audio, video, image, tactile)*

# Minimal schema – overview 2

- *ValidationInfo*

  validated

- *FundingInfo*

  projectTitle

  fundingType

- *MetadataInfo*

  metadataCreationDate

  source (for harvested metadata)

  harvestingDate

  originalMetadataLink

# Minimal schema – overview 3

- *PersonInfo*

  surname

  givenName

  CommunicationInfo

- *OrganizationInfo*

  organizationName

  CommunicationInfo

- *CommunicationInfo*

  email, position (title), organizationShortname, departmentName, adress, zipcode, city, country, region, telephone number, fax number, URL, affiliation

# IdentificationInfo

- *resourceTitle – the complete resource title*
- *pid* – persistent identifier pointing to the described resource
- *identifier* – unique identifier

# Example: IdentificationInfo

- SALDO
  - *resourceTitle*: Swedish Associative Thesaurus
  - *pid*: [handle]
  - *identifier*: saldo
- BLOGGMIX
  - *resourceTitle*: BLOGGMIX
  - *pid*: [handle]
  - *identifier*: bloggmix

# ContentInfo

- *description* – free text description

- *resourceType* – values: corpus; lexicalConceptualResource; languageDescription; technologyToolService

- *mediaType* – values: text; audio; video; image; tactile

# Example: ContentInfo

- SALDO
  - *description:* SALDO (Swedish Associative Thesaurus version 2) is an extensive lexical-semantic resource for modern Swedish written language.
  - *resourceType:* lexicalConceptualResource
  - *mediaType:* text
- BLOGGMIX
  - *description:* *BLOGGMIX is a dynamic corpus containing texts from Swedish blog sites.*
  - *resourceType:* corpus
  - *mediaType:* text

# DistributionInfo

- *availability* – free text

- *licenseInfo* – recommended values: GNU; CC; own; ELRA_END_USER; ELRA_VAR; ELRA_EVALUATION

- *distributionMedium* – recommended values: internetBrowsing; download; CD-ROM; DVD-R; bluRay; hardDisk; paperCopy; other)

# Example: DistributionInfo

- SALDO
  - *availability:* Available – restricted use
  - *licenseInfo:* CC-BY-SA 3.0; *LGPL 3.0*
  - *distributionMedium:* internetBrowsing; download
- BLOGGMIX
  - *availability:* Available – restricted use
  - *licenseInfo:* CC-BY-SA 3.0; *LGPL 3.0*
  - *distributionMedium:* internetBrowsing; download

# ValidationInfo

- *validated* – values: yes; no

# Example: ValidationInfo

- SALDO
  - *validated*: no
- BLOGGMIX
  - *validated*: no

# FundingInfo

- *projectTitle* – full title of the project that led to the creation of the resource

- *fundingType* – type of funding (e.g., EU, national, private, organisation funds, own funds, etc.)

# Example: FundingInfo

- SALDO
  - *projectTitle*: Språkbanken
  - *fundingType*: national, own funds
- BLOGGMIX
  - *projectTitle*: Språkbanken
  - *fundingType*: own funds

# MetadataInfo

- *metadataCreationDate* – for creation of metadata from scratch
- *source* – the catalogue from which the harvesting was made (CLARIN, OLAC, META, …)
- *harvestingDate* – date of harvesting of the metadata
- *originalMetadataLink* – link to the metadata of the original source

# Example: MetadataInfo

- SALDO
  - *metadataCreationDate*: N/A
  - *source*: N/A
  - *harvestingDate*: N/A
  - *originalMetadataLink*: N/A
- BLOGGMIX
  - *metadataCreationDate*: N/A
  - *source*: N/A
  - *harvestingDate*: N/A
  - *originalMetadataLink*: N/A

# PersonInfo

- *surname* – contact person's last name
- *givenName* – contact person's given name
- *CommunicationInfo* – how to reach the contact person for specified resource

# Example: PersonInfo

- SALDO
  - *surname:* Forsberg
  - *givenName:* Markus
  - *CommunicationInfo:* *markus.forsberg*@svenska.gu.se, SB (UGOT), Department of Swedish Language, Lennart Torstenssonsgatan 8, 405 30 Gothenburg, Sweden, Europe, tel.+46 (0)31 786 45 45, fax +46 (0)31-786 10 64, http://spraakbanken.gu.se/eng/personal/markus
- BLOGGMIX
  - *surname:* (same as above)
  - *givenName:* (same as above)
  - *CommunicationInfo:* (same as above)

# OrganizationInfo

- *organizationName* – name of the organization with the specified resource

- *CommunicationInfo* – how to reach the organization for specified resource

# Example: OrganizationInfo

- SALDO
  - *organizationName:* Språkbanken
  - *CommunicationInfo:* sb-info@svenska.gu.se, SB (UGOT), Department of Swedish Language, Lennart Torstenssonsgatan 8, 405 30 Gothenburg, Sweden, Europe, tel.+46 (0)31 786 00 00, fax +46 (0)31-786 10 64, http://spraakbanken.gu.se/
- BLOGGMIX
  - *organizationName:* (same as above)
  - *CommunicationInfo:* (same as above)

# Dependency: mediaType=text

- *LanguageInfo*
  - languageCoding
  - languageID
  - languageName
- *SizeInfo*
  - size
  - sizeUnit
- *AnnotationInfo*
  - annotationType
- *DomainInfo*
  - domain
- *FormatInfo*
  - mimeType
- *CharacterEncodingInfo*
  - characterEncoding

# LanguageInfo

- *languageCoding* – which standard used
- *languageId* – language identifier
- *languageName* – human understandable name

# Example: LanguageInfo

- SALDO
  - *languageCoding*: ISO 639-3
  - *languageId*: swe
  - *languageName*: Swedish
- BLOGGMIX
  - *languageCoding*: ISO 639-3
  - *languageId*: swe
  - *languageName*: Swedish

# SizeInfo

- *size* – numerical value

- *sizeUnit* – example units: words; tokens; bytes; sentences; texts

# Example: SizeInfo

- SALDO
  - *size:* *23.3*
  - *sizeUnit:* MB
- BLOGGMIX
  - *size:* *202 million*
  - *sizeUnit:* tokens

# AnnotationInfo

- *annotationType* – values: segmentation; alignment; structural annotation; lemmatization; stemming; PosTagging; bPosTagging

# Example: AnnotationInfo

- SALDO
  - *annotationType:* N/A
- BLOGGMIX
  - *annotationType:* segmentation; structural annotation; lemmatization; PosTagging; etc

# DomainInfo

- *domain* – indicates the application domain.

# Example: DomainInfo

- SALDO
  - *domain:* N/A
- BLOGGMIX
  - *domain:* N/A

# FormatInfo

- *mimeType* – the mime-type of the resource. Values taken form *IANA* (Internet Assigned Numbers Authority).

# Example: FormatInfo

- SALDO:
  - *mimeType: application/xml*
- BLOGGMIX:
  - *mimeType: application/xml*

# CharacterEncodingInfo

- *characterEncoding* – recommended values: ISO 8859-1; UTF-8; ISO 2022; etc.

# Example: CharacterEncodingInfo

- SALDO
  - *characterEncoding:* UTF-8
- BLOGGMIX
  - *characterEncoding:* UTF-8