Developing a Framework for Describing Relations among Language Resources

Penny Labropoulou[1], Christopher Cieri[2], Maria Gavrilidou[1]
[1]Institute for Language and Speech Processing/Athena Research Center, Athens, Greece
[2]University of Pennsylvania, Linguistic Data Consortium
E-mail: penny@ilsp.gr, ccieri@ldc.upenn.edu, maria@ilsp.gr

Abstract

In this paper, we study relations holding between language resources as implemented in activities concerned with their documentation. We envision the term "language resources" with an inclusive definition covering datasets (corpora, lexica, ontologies, grammars, etc.), tools (including web services, workflows, platforms etc.), related publications and documentation, specifications and guidelines. However, the scope of the paper is limited to relations holding for datasets and tools. The study focusses on the META-SHARE infrastructure and the Linguistic Data Consortium and takes into account the ISOcat DCR relations. Based on this study, we propose a taxonomy of relations, discuss their semantics and provide specifications for their use in order to cater for semantic interoperability. Issues of granularity, redundancy in codification, naming conventions and semantics of the relations are presented.

Keywords: language resource, metadata, relation, language technology, standardisation

1. Introduction – State of the art

Given the nearly universal recognition of the critical role of digital data in modern research and technology development, many communities are currently involved in data management, from various perspectives and with different objectives and goals: from (any form of) data archiving to research data publishing and to language data access and reuse.

The research data archiving perspective, for instance, focuses on the concept of publication as research product and uses it as their core entity, which is related to other entities, such as the author(s), the publisher, and recently, the related research data.

In the Language Resources (LR) and Human Language Technology (HLT) communities, on the other hand, the focus lies on the constellation of LRs: datasets (including corpora, lexicons, ontologies, grammars, etc.), but also tools, specifications, technical papers and documentation describing how data is created and processed (Bird and Simons, 2003), as well as the actors and processes that transform the various resources into new ones.

However different the perspective or the definition of 'data', 'content' or 'resource' in the various fields might be, what is common is the perception that discovering, accessing, documenting, reusing and interlinking data is crucial. To that end, the Linked Data initiative is viewed as best practice recommendation for exposing, sharing, and connecting pieces of data published on the Web (Berners-Lee, 2009). More specifically, Heath & Bizer (2011) state that "... *external RDF links* are fundamental for the Web of Data as they are the glue that connects data islands into a global, interconnected data space and as they enable applications to discover additional data sources in a *follow-your-nose* fashion".

There has been considerable, though not yet adequate, effort to standardize and/or map descriptions of the entities of the field of LRs (Broeder et al., 2009; Gavrilidou et al., 2011; Ahtaridis, Cieri and DiPersio, 2012, Mariani, et al., 2014). There has been less effort to date in identifying and characterizing the relations among LRs and the related entities (publications, provenance information, actors and process in transforming the resources etc.), although considered equally important (see, for instance, Chiarcos et al., 2012 and especially van Erp 2012, focusing on language data and metadata).

In this paper, we describe efforts to explore relations among LRs as implemented in relevant initiatives, focusing on the META-SHARE infrastructure (www.meta-share.eu) and the Linguistic Data Consortium (LDC, https://www.ldc.upenn.edu) but also the relations that have been included in the ISOcat DCR; based on this study, we propose a taxonomy of relations among LRs, discuss their semantics and provide specifications for their use.

We envision the term "language resources" with an inclusive definition covering datasets, tools related publications and documentation and the relations among these. We see this inclusive definition as critical to allowing the field to progress toward a state where 1) data creators can study existing methods as a pre-requisite to beginning their own work and can later monitor feedback on their products and their impact on the field, 2) tool developers can learn efficiently the goals and limits of data sets and can build upon previous treatments of them and 3) authors can discover and survey related work. As a first step, the focus of this paper lies on a subset of resources (i.e. datasets and tools/technologies) and the relations among them.

2. Overview of the treatment of LR relations in relevant initiatives

2.1 META-SHARE Relations

The central entity in META-SHARE is the LR, which is defined as encompassing **datasets** (corpora, lexical/conceptual resources and language descriptions) and **technologies** (tools/services) used for their processing. An important aspect of their description lies in their linking to other satellite entities, covering the LR's lifecycle from production to use:

- reference documents related to the LR (papers, reports, manuals, specifications, etc.),
- actors, i.e. persons and organizations involved in its creation and use (creators, distributors, etc.),
- related projects and activities (funding projects,

activities of usage, etc.) and
- accompanying licences.

Satellite entities are described only when the case arises, i.e. when linked to a specific LR.

Limiting the scope of the current paper to relations between LRs as defined above, we discuss the way these have been implemented in the META-SHARE metadata model (Gavrilidou et al, 2012), and the additional relations proposed by the users of META-SHARE, within the specific project but also within other projects that have deployed the model for the description of LRs.

A set of 9 relations were identified at the design phase for inclusion in the schema, covering relations:

- holding between resources of the same type (i.e. among datasets / tools): **derivedResource**, **originalSource**, **requiredLRs**
- connecting datasets with tools used for their processing and/or providing access to them: **accessTool**, **annotationTool**, **creationTool**, **evaluationTool**, **requiredSoftware**, **validationTool**.

This set was established after a survey of previous schemas and discussions with actors in the field, including participants in the satellite projects.

In these relations, LRs are connected to endogenous LRs (i.e. LRs included in the catalog) or exogenous LRs (described at other web sites and catalogs). A free text field is provided for the encoding of the exogenous resource, so that metadata creators can enter a URL, name of the resource, or identifier (e.g. "http://www.statmt.org/moses/", "MOSES", or "MOSES MT").

2.2 META-SHARE+ Relations

To accommodate needs that had not been covered in this initial set and to facilitate the proposal of additional relations by users, an additional component, **relationInfo,** has been included in the META-SHARE model. The component includes two elements, the **relationType** to be used for naming the relation and the **relatedResource** which specifies the target resource.

As of the time of writing, users have proposed a set of 11 values, which have been used 332 times for 39 resources covering the following cases:

- relations between LRs of the same type: **alignedWith**, **isSpokenVersionOf**, **linkBetween**, **partOf**, **hasPart**, **isExtensionOf**, **derivedFromSameSource**, **derivedFrom**, **source corpora**, **the source corpus**
- relations between datasets and processing tools: **required software**.

Taking a closer look at the proposed values and the way these have been used in the metadata entries, we see that:

- two relations, **source corpora & the source corpus**, bear a resemblance to the **originalSource** of the proposed set. One of these cases (PELCRA Word Aligned Corpora) refers to the superset corpora from which the resource has been created. The other case (Copenhagen Dependency Treebank) provides the link to the morphosyntactically annotated version of the corpus, i.e. the previous stage in the processing chain.
- the relation **alignedWith** is used to link together the monolingual parts of the META-NORD ACQUIS and Sofie multilingual parallel corpora
- the relations **derivedFrom** and **derivedFromSameSource**[1] encode the fact that the two Norwegian wordnets are based on the Danish wordnet, i.e. this resource has been used as a model for the construction of the two new resources
- **partOf** and its inverse relation **hasPart** are deployed for cases where a resource can be split into subsets on the basis of various features: at the language dimension, for instance, they are used for the monolingual parts of parallel corpora (e.g. META-NORD Danish Sofie Parallel Treebank); also, a corpus of recordings of various participants split into subsets per participant (e.g. Tactile Reading SB is reading braille)
- **linkBetween** is used twice in the same entry (e.g. Finnish – Danish linked wordnets) to bring together two monolingual wordnets combined into a bilingual one – a more specific case of a part-whole relation
- **isExtensionOf** is used for a special kind of part-whole relation again, where the resource considered as "part" is enriched/extended at a later stage and results in a new resource (cf. LEXIS Computational Lexicon which is the continuation of the Greek PAROLE and SIMPLE lexica)
- as its name reveals, **isSpokenVersionOf** relates two versions of the same resource, namely a text resource and the recording thereof (cf. Database of Bulgarian speech recordings)
- finally, the **required software** relation is used for the same purposes as the originally proposed **requiredSoftware** and its use is probably due to some misunderstanding of the schema.

Another project that deploys the META-SHARE schema is QTLaunchPad (QTLP, http://www.qt21.eu/launchpad/). One of its aims is to investigate the potential of automatic discovery and processing of LRs with web services in the context of Machine Translation; the discovery procedure is based on the LRs' metadata descriptions.

The **relationInfo** component has been used by QTLP to encode 5 new relations:

- **isAnnotationOf** is used, of course, for the annotated versions of raw corpora; in fact, this relation is automatically added to the metadata of the resources that result from the application of the web services on LRS included in the repository
- **isSimilarTo** associates processed versions of the same resource resulting from the operation of different tools, e.g. two versions of the same corpus annotated at the same level but with a different tool (e.g. JRC-Acquis subcorpus EN-DE with HunAlign and with Vanilla)
- **isRelatedTo** is a general term covering two distinct subcases: (a) aligned versions of the same corpus but at different levels (word vs. sentence alignment), and (b) a parallel multilingual corpus which is composed

[1] The latter relation could be considered redundant since it can be indirectly deduced from the former.

of different bilingual subsets

- **isSubsetOf** and **isSampleOf** belong both to the part-whole type; what differentiates them is that the former is used for monolingual parts of parallel corpora (e.g. JRC-Acquis EN-DE **isSubsetOf** JRC-Acquis 3.0) while the latter is used for small parts of a resource which are available for free in a demo-like fashion.

2.3 LDC Corpus Relations

In addition to the documentation supplied with each corpus it publishes, LDC also creates a description to include in its Catalog. The intended readers are of course, potential users. The Catalog description typically mentions related corpora. The goal of the Catalog description is not to enumerate every possible relation among corpora but to mention those deemed relevant to potential users. By surveying those descriptions, we were able to identify numerous relations and relation types that exist among LDC corpora. We lack similar information for relations involving non-LDC corpora. We should also make clear that these are almost certainly a subset of all relevant relations among LDC data resources and that they say nothing about relations between data sets on the one hand and tools, specifications, or technical papers on the other.

Of the 574 corpora LDC had published at the time of writing, 337 have Catalog descriptions that mention other LDC corpora. If we take this as representative of the field, it means that more than half of all data sets are related to one or more other data sets. This fact alone should make it clear why the study of LR relations is important to the field.

By manually reviewing the Catalog descriptions we were able to identify the following relation types either because they were mentioned within the description or because they became apparent to the reviewer upon further inspection.

The various TIMIT corpora stand in several different relations to each other. The original corpus TIMIT (LDC93S1W) contains recordings of multiple subjects reading phonetically rich sentences into a close talking microphone. The prompts and audio are aligned at the sentence, word and phoneme level and the corpus includes metadata on the readers. FFMTIMIT (LDC96S32) contains recordings of the same session but uses a far-field microphone. The other TIMIT-derived corpora differ from FFMTIMIT in that they derive not directly from the original source but rather from the published recordings. That is FFMTIMIT records the original source through a different audio channel while the other re-records it through a second channel. CTIMIT (LDC96S30) transfers TIMIT though multiple cellular telephone circuits while HTIMIT (LDC98S67) uses different handsets, NTIMT (LDC93S2) uses the NYNEX telephone network and WTIMIT (LDC2010S02) uses a wide-band mobile network. Finally, to reduce the effects of channel variation over time, STC_TIMIT transmits a subset of TIMIT over the telephone network in a single call. From these few examples, we see that some corpus may **contain, sample** (i.e. contain a subset), **re-record**, and **re-encode** another corpus. We also see an example of **part-whole** relationships. By 1996, LDC had received and released another recording of the original TIMIT sessions via a secondary far-field microphone called FFMTIMIT. We may view TIMIT and FFMTIMIT then as two parts of a whole that was never published as such. Finally, given that our purpose here is resource discovery and complete description, we believe it is wise to abstract from the numerous details of audio sampling, encoding and simply mark when a corpus is related to another such that the original signal differs.

We can further explore the part-whole relationship types with the ATIS0 corpora created to support the development of an Air Travel Information System. ATIS0 was distributed in three parts and as a complete set. The first part ATIS0 Pilot (LDC93S4B) contains 912 spontaneous utterances from 36 speakers collected via a Wizard of Oz protocol in which subjects interacted with the system to identify flight options for a given itinerary. Subjects' speech was recorded via close-talking and desktop microphones. In ATIS0 Read (LDC93S4B-2), 20 of the original 36 speakers read a total of 478 versions of the utterances from the Pilot corpus. In ATIS0 SD (LDC93S4B-3) ten of the same speakers recorded speaker dependent material in the ATIS domain. ATIS0 Complete (LDC93S4A) **contains** all three of these corpora. Each of the three is **part-of** the complete set and a **part-with** the other two. These part-whole relationship are common, affecting for example the 4 Resource Management corpora (e.g. LDC93S3A, Complete Resource Management corpus 2.0), 6 CSR corpora (e.g. LDC93S6A, CSR-I (WSJ0) Complete), 4 TIPSTER corpora (e.g. LDC93T3A, TIPSTER Complete), 4 ATC0 corpora (e.g. LDC94S14A, Complete ATC0) and 4 UN Parallel Text corpora (LDC94T4A, Complete Parallel Text) among many others.

The Switchboard corpora reveal different relationship types. During the intense use and re-annotation the original Switchboard enjoyed, users identified and fixed problems with the file inventory and speaker attribution and added metadata and annotations. The resulting version, Switchboard-1 Release 2 (LDC97S62) **replaces** or **supersedes** the original Switchboard-1 (LDC93S7) in the sense that it is received wisdom to use the revised version. Treebank-2 (LDC95T7) bears a similar relationship to the original Treebank release (LDC94T4B).

Another very common relation type is found among the CALLHOME corpora. For example CALLHOME Mandarin Chinese Transcripts (LDC96T16) **annotates** CALLHOME Mandarin Chinese Speech (LDC96S34) but does not contain it. The speech and transcripts are published separately. The CALLHOME Lexicons presumably bear a different relation type, presumably **derived_from** the speech and transcripts. These relationship recur in the other triples of CALLHOME corpora in American English, (e.g. LDC97S42), Egyptian Arabic (LDC97S45), German (LDC97S43), Japanese (LDC96S37) Mandarin (LDC96S34) and Spanish (LDC96S35). There is an additional relationship among these in that all of the CALLHOME corpora were created to fulfill a similar purpose. However, the **part_with**

relation seems imperfect as these corpora were created at different times by different teams. The aspect that one would want to capture is that the corpora are **in_series_with** each other and rely upon similar specifications. Such series are common including 15 CALLFRIEND corpora (e.g. LDC96S50 in Farsi) and 12 JEIDA corpora (e.g. LDC96S64, JEIDA/ JCSD-Channel 0 Complete) among many others.

A similar but not identical relationship connects two MUC corpora. Message Understanding Conference (MUC) 6 Additional News Text (LDC2003T13) **continues** Message Understanding Conference (MUC) 6 (LDC96T10) but was created several years later. The difference between the MUC corpora on the one hand and the CALLHOME, CALLFRIEND and JEIDA corpora on the other is that the latter were always intended to comprise a series whereas as the MUC continuation was conceived and developed much later than the original. At a somewhat greater distance, the VAHA (LDC96S41) corpus was **inspired_by** the MACROPHONE (LDC94S21) corpus though created independently.

We also need some kind of identity relationship for cases in which a single corpus is known by different names at different times or different data centers. For example Message Understanding Conference (MUC) 6 Additional News Text (LDC96T10) **equals** or **renames** MUC VI Text Collection the original name under which identical data was published.

2.4 Relations in ISOcat

The ISOcat Data Category Registry (DCR) is an ISO 12620:2009 compliant registry for elaborate specifications of data categories (ISO 12620, 2009). It has been set up to serve semantic interoperability through the registration of *elements* ("data categories"), which refer to widely used concepts in the linguistics domain; users can then link their own elements to them (or add new ones according to the ISO 12620 requirements), thus achieving common terminology. A thematic area on metadata is included.

In the DCR there is no distinction between elements describing the properties of a resource and those denoting a relation between two resources. For the purposes of this paper, we have gone through the contents of the metadata area and tried to identify relations on the basis of the definitions and examples included for each data category[2]. The identified relations fall under the general categories:

- the largest set of elements associates data resources with tools used for their creation, processing, management and usage: **accessTool, analysisTool, annotationTool, archivingTool, creationTool, deploymentTool, derivationTool, derivationWorkflow, displayTool, editingTool, elicitationSoftware, queryTool, recordingPlatformSoftware**[3]
- one more element, **runningEnvironment**, can be used both for tools and data resources
- the element **originalSource** is used for the description of LRs produced on the basis of other LRs
- finally, the **relationType** is meant as a generic element that allows users to name the relation.

3. Analysis of the findings

The increasing addition of information on relations, mainly as regards documents describing the resources, but also between datasets and tools that have been or can be used for their processing, shows that LR providers consider them important for the documentation and promotion of their resources. To maximise the benefits from this knowledge, standardization of the relation values is deemed indispensable. Aiming at this target, we proceeded with:

- assessment/comparison and contrast of the above relations, in order to find the commonalities and differences between them, clarify their semantics, eliminate possible duplicates and treat differences attested at the level of granularity
- classification and clustering of the relations into broad categories
- formalization of these categories into a taxonomy of relations, each one accompanied with a proposed naming convention, definition and specifications of use. The proposed taxonomy is compatible with all reviewed approaches, catering for interoperability.

3.1 Assessment of relations proposed

The comparative study revealed commonalities between the approaches of META-SHARE, META-SHARE+, LDC and ISOcat, which mostly concern the types of relations these initiatives selected to document. The differences observed are connected to the level of granularity (e.g. the relation connecting two datasets **isVersionOf** versus **isAnnotatedVersionOf** or **isTaggedVersionOf**), the naming conventions (e.g. **annotates** versus **isAnnotatedVersionOf**) and the focus put on the source or the target dataset/tool described (e.g. the relations **derivedResource** and **OriginalSource** essentially describe the same relation, i.e. the relation holding between a resource which, through some process or transformation, produces another resource, but the first focuses on the outcome and the second on the source).

Here, we distinguish four broad classes depending on the LR types connected via the relations, namely relations ***between datasets***, ***between tools***[4], ***between datasets and tools***, and ***between any type of resources***. Within each class, the relations are grouped together according to the type of the relation described.

3.1.1 Relations between datasets

This class groups together relations that connect data

[2] The study ended in February 2014; the DCR is constantly being enriched with new elements and modifications of definitions, so the findings of this study reflect the state of the DCR at this time interval. Moreover, definitions are not always clear, so the list of relations presented here may not be exhaustive.

[3] META-SHARE intentionally deployed existing data categories when available for the sake of semantic interoperability, which explains the similarity in the names of the elements, especially those of the first set of relations.

[4] The term "tool" is meant to cover tools, web services, workflows, platforms, and, in general, any kind of s/w.

resources based on the following features:

- **part whole relation**

This type of relation is one of the most widely used and refers to the case of an LR that is (or includes) a subset of another; the two LRs are provided both together and as separate resources; this is the case, for example, of monolingual parts of parallel corpora, the entries of the syntactic level of a lexicon, the subtitles or audio part of a video, a sample provided for demo purposes, resources published independently and as series, etc. This relation is manifested by names such as **partOf**, **source corpora, isSubsetOf**, **isSampleOf**.

- **transformation**

The resources connected constitute two stages in the processing of the same entity; i.e. one is the outcome of a transformation on the other. This broad class accommodates different processing levels and/or formats; i.e. cases such as the relation between a terminological list and the corpus from which it was extracted, between an audio corpus and its re-recording through a different audio channel, etc. Names of relations are **derivedFrom, isSpokenVersionOf, originalSource, re-encodes** etc.

- **combination**

This relation describes the connection of resources that combine together to form a third resource, i.e. the relation holding between the two (or more) parts rather than between the parts and the whole. Such cases are, for instance, two monolingual corpora aligned to constitute together a parallel corpus, or two levels of a lexicon (e.g. morphological and semantic) included in a three level computational lexicon. Such relations are named, for example, **linkBeween, part_with**.

3.1.2 Relations between tools

This class comprises two relation types:

- **prerequisites**

This relation codifies the requirements set by a tool as regards another tool or environment.

- **evaluation**

In this case the relation connects a tool with the software used for its evaluation.

Names used are, indicatively, **required_software, isEvaluatedBy.**

3.1.3 Relations between datasets and tools

The majority of relations fall under this class, which subsumes three broad sub-classes.

- **creation tools**

The relation coded here is the relation between a dataset and its creation tool, e.g. web crawler, OCR tool, term extractor, recording s/w etc., as attested by the relations **creationTool, elicitationTool, derivationTool**, etc.

- **processing tools**

The relations that belong to this class connect datasets with the tools that have annotated (at any level of annotation), analysed, edited or validated them. Relations proposed are **annotationTool, validationTool, analysisTool,** etc.

- **management tools**

This class collects relations of datasets with tools used for accessing, archiving, displaying or querying them. The initiatives overviewed have proposed names such as **queryTool, archivingTool, accessTool** and **displayTool.**

3.1.4 Relations between any type of resources

These relations are constrained by the fact that they connect same types of LRs and belong to three broad categories.

- **sameness**

This relation connects resources that are available with different names but have identical content (in the case of datasets) or code (in the case of tools), and resources that have for some reason changed name[5].

- **similarity of specifications or principles of creation**

This relation connects LRs which present a greater or smaller degree of similarity, in the sense that they adhere to the same principles or specifications, were created with a similar purpose or were derived from the same source. Similarity here is to be interpreted qualitatively and not as a quantitatively calculated measure. Examples include WordNets for different languages, speech corpora following the same recording specifications, etc.

- **Versioning**

Relations encoding the extension of a resource as regards its size, addition of annotation to a dataset, correction of the content of a dataset, debugging of a tool etc., belong here. Note that in the previous section we list relations between tools and datasets, while here we classify relations between the initial resource and the updated one. These resources can be viewed as two stages in the evolution of the resource, connected by relations such as **isAnnotationOf**, **version**, **isVersionOf**, **replaces**, etc.

3.2 Proposal for the codification of relations

The relations discussed above are presented in tabular form in the Appendix. The relations proposed by the three initiatives were compared, grouped together according to their semantics, their intended use and the resources they apply to and, finally, classified into the above discussed classes. For each relation we give a definition, an example or comment, names used by each initiative and finally a proposed name aiming at transparent semantics.

Some of the issues we took into consideration for the construction of the proposed taxonomy are:

- ***Naming specifications***: we opted for the use of verbal expressions, for two reasons: (a) the arguments of verbs are more transparent than those of deverbal nouns; thus, **isAnnotatedBy** is a semantically transparent expression that connects resource A to resource B, and (b) verbs also clearly specify the direction of the relation; e.g. resource A (a dataset) **isAnnotatedBy** resource B (a tool), whereas this would not be evident if the name **annotationTool** was used. Similarly, the **partOf** relation has been renamed **isPartOf, hasPart** etc.
- ***inverse relations***: the inclusion of pairs of inverse

[5] This issue would clearly benefit from the establishment of the ISLRN (Choukri et al, 2012).

relations (such as **annotates** and **isAnnotatedBy**, **hasOutcome** and **hasOriginalSource**) makes the model more expressive but increases its redundancy. Even in the case of only adding one relation, an intelligent search allows the discovery of the other; e.g. if resource A is described as **hasOutcome** resource B, a guesser would find that resource B **hasOriginalSource** resource A. Still, the proposal should cater for all relations given that we cannot predict which one of the two resources will be described in a catalog or repository, the outcome or the original source. Thus, inverse relations are included in the proposal but users are advised to encode only one in cases where both resources are included in a catalog.

- ***Level of granularity***: we opted for the middle solution between very broad relations and too fine-grained ones. For each class, we propose a set of broad/top categories and, depending on the class and its requirements, a set of finer relations. The proposed names try to subsume too fined-grained relations. Still, users that wish to make finer distinctions can do so, provided that they adhere to the same naming conventions. The use of qualifying adjectives (e.g. **isUpdatedVersionOf**) or adverbs (e.g. **containsPartially**) where possible is recommended.
- ***Target resource***: the name shows clearly the direction between the two arguments; argument A is the resource being described and resource B is the target resource. If the target resource is also included in the same catalog of resource descriptions, they can be linked via the id mechanism of this catalog. Otherwise, reference to an exogenous resource is difficult. In the initiatives we have studied, users were hesitant between a url link (the page describing or containing the resource) and the name of the resource or both.

4. Conclusions and future tasks

In this paper we have proposed a taxonomy of relations between LRs, to be included in their metadata documentation or catalog description. Future plans include:

- application of the taxonomy to the resources included in META-SHARE and LDC
- conversion of the META-SHARE metadata model into RDF in order to better accommodate and encode the relations and establishment of a mechanism for extending and updating the taxonomy
- study of the possibility of (semi-)automatically discovering relations from the LR documentation, academic papers, free text descriptions in metadata, etc.
- extension of the taxonomy to include relations with publications and specifications.

5. Acknowledgements

This work has been partially supported by funding from the EU FP7-ICT QTLaunchPad project (grant agreement no. 296347) and the Greek Operational Programme "Competitiveness and Entrepreneurship" (OPCE II) CLARIN-EL project (MIS code 441451).

6. References

Ahtaridis, E., C. Cieri and D. DiPersio (2012), LDC Language Resource Papers: Building a Bibliographic Database, LREC 2012 Istanbul, Turkey, May 23-25.

Berners-Lee, T. 2009. Linked Data. In *Design Issues*. World Wide Web Consortium. www.w3.org/DesignIssues/LinkedData.html.

Bird, S. & Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, 79:557–82

Broeder, D., Gaiffe, B., Gavrilidou, M., Hinrichs, E., Lemnitzer, L., Van Uytvanck, D., Witt, A., Wittenburg, P. (eds.) (2009). CLARIN Deliverable D2.4 - Metadata Infrastructure for Language Resources and Technology http://www.clarin.eu/sites/default/files/wg2-4-metadata-doc-v5.pdf

Chiarcos, C., Nordhoff, S., Hellman, S. (2012). Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata. Springer. DOI 10.1007/978-3-642-28249-2

Choukri, K., Arranz, V., Hamon, O. and Park, J. (2012). Using the International Standard Language Resource Number: Practical and Technical Aspects. Proceedings of the Eighth International Conference on Language Resources (pp. 50 - 54).

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerq, T., Francopoulo, G., Arranz,. V., Mapelli, V. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. Proceedings of the Eighth International Conference on Language Resources (pp. 1090 - 1097). ELRA

Gavrilidou, M., Labropoulou, P., Piperidis, S., Speranza, M., Monachini, M., Arranz, V., Francopoulo, G. (2011). META-NET Deliverable D7.2.1 - Specification of Metadata-Based Descriptions for Language Resources and Technologies.

Heath, T. and C. Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology 1(1). Morgan & Claypool. http://linkeddatabook.com/editions/1.0/.

ISO 12620 (2009). Terminology and other language and content resources - Specification of data categories and management of a Data Category Registry for language resources, International Organization for Standardization.

Mariani, Joseph, Christopher Cieri, Gil Francopoulo, Patrick Paroubek, Marine Delaborde (2014) Facing the Identification Problem in Language-Related Scientific Data Analysis In Proceedings of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, 26-31 May 2014.

van Erp, M. (2012). Reusing Linguistic Resources: Tasks and goals for a Linked Data approach. In Chiarcos, C. & Nordhoff, S. & Hellmann, S. (2012), *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata.* Springer. DOI 10.1007/978-3-642-28249-2, pp. 57-64.

Definition	Example / comment	META-SHARE	MS+	LDC	ISOCat	Renaming proposal
Cases where DATASET B is the outcome of some type of processing on DATASET A						
LR A has been used as the basis / initial / source material from which LR B was created /extracted	terminological list as the result of term extraction process in a corpus	DerivedResource				hasOutcome
LR A is the product / outcome / of LR B (the original source)		OriginalSource	derivedFrom, the source corpus	derived_from	originalSource	hasOriginalSource
LR B re-records LR A	Recording through a different audio channel			re-records		isPartWith
LR B re-encodes LR A				re-encodes		hasOriginalSource
LR A is the text that informants uttered in the recording of LR B			isSpokenVersionOf			hasOriginalSource
Cases where RESOURCE B is in some way similar to RESOURCE A						
LR B is the new / alternative name for A, while the content is identical	the ISLRN can solve the problem			is_equal_to /equals / renames		isSameAs
LR B is similar to A as regards creation specifications, purpose, source material etc.; created as part of a series/set of similar resources	the classic example is that of WordNets		derivedFromSameSource	in_series_with		isSimilarWith
LR B has been inspired by A but without strictly adhering to the same principles; not considered as a series/set			basic theory and methodology	inspired_by		isSimilarWith
LRs A and B are annotated at the same level but with different tools			isSimilarTo			isSimilarWith
Cases where one DATASET is part of another DATASET						
LR A is part of LR B	monolingual parts of parallel corpus, syntactic level of a lexicon, subtitles of a video, demo sample		partOf, source corpora, isPartOf, isSubsetOf, isSampleOf	part_of		isPartOf
LR A contains LR B	inverse relation		hasPart	contains, is_sample_of		hasPart
Cases where one DATASET is combined with another DATASET						
LRs A and B are parts of LR C				part_with		isPartWith
LR A is aligned with LR B	two monolingual WordNets / corpora aligned to produce a bilingual resource		alignedWith, linkBetween			isCombinedWith
Cases where a TOOL B is used for creating the DATASET A						
LR A was created with tool B	web crawler, term extractor	creationTool			creationTool, derivationTool,	isCreatedBy

Definition	Example / comment	META-SHARE	MS+	LDC	ISOCat	Renaming proposal
					derivationWorkflow	
LR A was elicited with s/w B					elicitationSoftware	isElicitedBy
LR A was recorded with tool B					recordingPlatformSoftware	isRecordedBy
Cases where a TOOL B is used for accessing/managing the DATASET A						
LR A can be accessed by tool B	corpus workbench, s/w for lexicon access	accessTool			accessTool	isAccessedBy
LR A can be queried by tool B	a corpus application with an interface for corpus query				queryTool	isQueriedBy
LR A is archived by tool B					archivingTool	isArchivedBy
LR A is displayed / visualized by tool B	Incl. visualization tools				displayTool	isDisplayedBy
Cases where a TOOL B is used for processing the DATASET A						
LR A was annotated by tool B		annotationTool		annotates	annotationTool	isAnnotatedBy
LR A was edited by tool B					editingTool	isEditedBy
LR A was analysed by tool B	statistical tools				analysisTool	isAnalysedBy
LR A was validated by tool B		validationTool				isValidatedBy
Cases where a RESOURCE is needed for the operation of TOOLS						
LR B is required for the operation of tool A	grammar for a parser, list of stop words	requiredLRs				requiresLR
S/w B is required for running tool A		requiredSoftware	required software		runningEnvironment	requiresSoftware
Tool A was evaluated by tool/metric/package B		evaluationTool				isEvaluatedBy
Relations connecting RESOURCES of the same type						
LR B continues LR A				continues		isContinuationOf
LR B is an extension in size, corrections of content, validation, debugging (for tools) of LR A			isExtensionOf, isUpdatedVersionOf, isAnnotationOf		version	isVersionOf *[possibly with adj., isAnnotatedVersionOf, isUpdatedVersionOf etc.]*
LRs A and B are annotated at the same level but with different granularity			isRelatedTo			isSimilarWith
LR B replaces or supersedes LR A				replaces		replaces