VIRTUAL MOBILITY (VM) GRANT REPORT TEMPLATE

This report is submitted by the VM grantee to VNS Manager, who will coordinate the approval on behalf of the Action MC.

Action number: 18209

VM grant title: Dictionary Metadata

VM grant start and end date: 06/09/2021 **to** 15/10/2021

Grantee name: Dr. David Lindemann

Description of the outcomes and achieved outputs (including any specific Action objective and deliverables, or publications resulting from the Virtual Mobility).

Main findings (more detailed in working document, and summary document):

The objective of this VM grant was to prepare the ground for the integration of metadata for printed and electronic dictionaries in LexBib, a digital bibliography and Knowledge Graph project for the domain of Lexicography and Dictionary Research, which currently stores metadata for lexicography-related publications.

To this end, we worked on the definition of a Dictionary Metadata (DM) model, re-using and extending relevant vocabularies. Our main sources are:

- the META-SHARE ontology (MS-OWL), an RDF vocabulary for the description of Language Resources, including Lexical resources (LR), until today mainly used for Natural Language Processing and Language Technology LR;

- the LexVoc vocabulary for lexicographic terms;

- the FRBR model and BIBO ontology for bibliographic citations.

We have determined the core classes of the new DM model, defining correspondences between FRBR and MetaShare. We have concluded that frbr:Expression is equivalent to ms:LexicalConceptualResource and frbr:Manifestation to ms:DatasetDistribution.

Combining the two approaches together may pose problems, as library catalogues include records for manifestations/distributions where all metadata categories are included in a flat list, while catalogues of datasets consist of records for expressions/resources, under which manifestations/distributions are represented, each with their own set of properties. On the other hand, the two-level structure of the datasets approach is well suited for our model, given that there are properties, such as type of dictionary, microstructure contents, etc., that describe the resource, irrespective of its physical representation(s). At the same time, dictionary metadata as needed for citation (BIBO) describe the manifestation, irrespective of whether other manifestations embody the same lexicographical expression.

MS-OWL includes one more level for the description of Language resources, based on the notion of "media type", the physical medium modality (i.e. text, audio, video, image) in which a resource or resource part is available. Multimedia dictionaries can also be described in a simple way with a limited set of metadata categories: in the simple approach, audio and video are not represented as autonomous MediaParts, but described in the text part using ms:ExtratextualInformation. We decided to keep the level as is, to allow support for complex cases as well, and to keep compliance with MetaShare on a high level, despite using only MediaPart 'text' for LexBib.

We have organized specific properties and values in a new structure. Parts of the LexVoc vocabulary, developed by the grantee, will be re-used, as well as properties defined in the ontology underlying LexBib. Properties describing dictionary content (attached at resource level) point to items in a range defined according to LexVoc top-level concept or "facet". For example, the "dictionary scope" property points to items defined as narrowers of top-level concept (facet) "dictionary scope".

We have discussed on mapping properties appearing in the MS-OWL, FRBR and BIBO entity schemata to each other, and decided on the class to attach them, i.e. domain Resource (expression) vs. Distribution (manifestation). For most properties, this was straightforward. For properties used for bibliographic citations, though, the two-level structure poses problems. We have discussed that in more detail, looking at identifier, distribution medium, format and access URL, that are potential sources of conflict. We have also proposed mappings of controlled vocabulary terms, both in MS-OWL and LexVoc.

We have created a set of items modeled according to our new DM model in LexBib wikibase.

We plan dissemination and feedback request actions in the frameworks of ELEXIS project, and Nexus Linguarum, and in the Lexicographic community. Regarding the latter, we plan to read a dedicated full paper at Euralex 2022 conference.

Description of the benefits to the COST Action Strategy (what and how).

This VM has enabled a knowledge transfer regarding RDF modeling and interoperability with state-of-the-art standard vocabularies from the Institute for Language and Speech Processing (ILSP) towards the grantee, who is currently engaged as postdoc researcher in the framework of Elexis H2020 research infrastructure project, in charge of developing and maintaining a Linked Database of bibliographical data, LexBib, and a vocabulary of lexicographic terminology, LexVoc. This VM has skilled him for designing both resources in a way that (1) general principles in ontology engineering, (2) interoperability with META-SHARE, a widely used vocabulary for the representation of Language Resource metadata, is guaranteed, and it has (3) established or strengthened existing communication channels from the grantee to members of relevant communities, such as those maintaining Ontolex and MetaShare.

This VM has fostered collaboration and synergies between researchers working in different disciplines (Lexicography, Libraries, Computational Linguistics, Metadata). More specifically, the grantee has been able to contribute viewpoints from the Lexicography perspective, on the one hand, and from the Library Science perspective, on the other. In both cases he has been supported by members of the Lexicography group in Hildesheim (Prof. Heid), and by Christiane Klaes, head of Metadata Department at Braunschweig TU Library. These viewpoints have been influencing the data model underlying LexBib research article bibliographical data, which will now be extended towards lexical/conceptual resources using the Dictionary Metadata model developed during the VM; in our working documents, we have spotted several aspects of that model that will also influence the ongoing discussion about a further development of the MetaShare ontology. In that sense, the benefit of this VM's outcomes is reciprocal.

The VM falls in the scope of Nexus Linguarum WG1 - Task 1.1 objectives (LLOD modelling). Through our interdisciplinary work, we have laid the foundations for a model for Dictionary Metadata, deploying Linked Open Data principles, by re-using and interlinking existing vocabularies. The model will be used in the creation of a catalogue of dictionaries, published as a LOD resource contributing to the LexBib knowledge graph, bringing together resources (dictionaries), publications, and agents in the field of lexicography. A summary of methodology and findings of this VM will be added to the Working Group 1 deliverable and a more detailed presentation will be shared with the Nexus Linguarum and ELEXIS communities, in order to get feedback and attract further collaboration.

Description of the virtual collaboration (including constructive reflection on activities undertaken, identified successful practices and lessons learned).

Collaboration has been carried out using telcos and cloud-based workspaces of different kind.

Despite being a Virtual Mobility, and not a traditional in-situ STSM, the action has allowed the grantee and the host to collaborate in an efficient way. Time schedules were agreed with similar regularity as they would have been set up in a STSM. Weekly telcos were held based on a structured agenda, and with detailed minutes (see separate document). All sessions have been recorded, and that has turned out very useful: telcos were not slowed down by minute taking, because that task has been done by the grantee after telcos, using the recording. In between telcos, collaborative work was done at any time sharing the same workspace using google docs, along with spontaneous and more informal exchange through google doc comments and email. The grantee prepared work in google spreadsheets, while linking it to different textual working documents, and to data sources on line.

For Linked Data modeling examples, we used the LexBib wikibase instance, which features instant data exhibitions, database queries, and graph visualisations on line, so that also data-modeling examples could be edited live while kept in the cloud. We have noted that this particular feature of wikibase as database solution eased our work, since other state-of-the art RDF database and client tools we have access to require non-trivial installation and maintenance processes for cloud-based working, hardly to put in practice during the lifetime of this VM.

We want to point out that the goal and the working agenda of the VM had already been agreed on in detail before starting the VM. In particular, we had agreed on concrete examples to test our models, so that we could soon start to discuss details.

Third-party collaborators joined the discussion on telcos, and/or via the virtual workspaces. Here, we suspect a weak point of VM vs. STSM: In six weeks time, it was possible to establish intense collaboration between the grantee and the host, but attracting more external collaboration than those external contributions that were foreseen from before the VM, and awaiting any feedback from larger communities, is hardly achievable. But, of course, it will be in the very near future. We will go on having weekly telcos, in order to collect and discuss that feedback, and prepare a more detailed report, as well as a documentation for our DM model, and a Euralex conference paper.

In general, we conclude that the VM as action format is efficient, since using tools for collaborative work can bridge the disadvantage of physical distance. In our case, regarding collaboration, we do not see significant losses compared to an in-situ collaboration of only six weeks duration, mainly because our object of study is at all digital, and best practices of collaborative work involve the use of the described cloud-based collaborative workspace platforms in any case, at least if additional collaborators take part virtually.