Language ~~Coding —~~ coding — A ~~Framework~~framework for ~~Language Varieties~~language varieties — Part 3:
Application of the framework

Identification et description des variétés de langues — Partie 3: Exigences et recommandations pour la mise en œuvre

FDIS stage

iTeh Standards
(https://standards.iteh.ai)
Document Preview

ISO/FDIS 21636-3
https://standards.iteh.ai/catalog/standards/iso/b814ca2c-7cfc-465f-a5ff-b3abb6acc223/iso-fdis-21636-3

                      © ISO 2023 – All rights reserved

Edited DIS - MUST BE USED FOR FINAL DRAFT

Contents

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documentsdocument should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawnISO draws attention to the possibility that some of the elementsimplementation of this document may beinvolve the subjectuse of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction or on the ISO list of patent declarations received (see), or both.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 37, *Language and terminology*, Subcommittee SC 2, *Terminology workflow and language coding*.

A list of all parts of the ISO 21636 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

An increasing amount of digital language resources (LRs) are being created (including via ~~retrodigitization~~retro-digitization), archived, processed and analysed. Within this context, the detailed and exact characterization of language varieties present in a given language use event is quickly gaining importance. Here, language use includes all modalities such as written, spoken, or signed, and also new forms of language use supported by digital technology (in social media and similar forms of digital communication). Such modalities demonstrate one way in which languages vary internally. Others include, for instance, familiar regional (dialectal) and social variation.

In the past, a primary goal of working with LRs was the archiving and preservation of LRs. However, new goals have now emerged and are still emerging:

— Institutions and individuals need to exchange metadata (i.e. bibliographic description data and other secondary information) for making the information on existing LRs widely available in a harmonized form.

— Researchers are identifying primary data (i.e. the LRs themselves) for various research purposes, including research on linguistic variation.

— Researchers and developers need LRs for the development of more advanced language technologies (LTs) and for testing purposes, because LTs, in particular those concerning speech recognition and language analysis, are entering more dimensions of human communication.

In order to achieve the above-mentioned goals and purposes, along with others not outlined in the ISO 21636 series, a standardized set of ~~metadata~~metadata for the identification of language varieties is important for guaranteeing the frictionless exchange of secondary information. Well-organized metadata also help to indicate the degree of interoperability (equalling re-usability and re-purposability of LRs), and the applicability of LTs to different situations or LRs over time. These metadata are applicable in eBusiness, eHealth, eGovernment, eInclusion, eLearning, smart environments, ambient assisted living (AAL), and virtually all other information-rich applications which depend on information about LRs. A clear metadata approach is also a prerequisite for the durability of LR archiving (in particular in the case of cultural heritage and scientific research data).

~~The International Standard~~ ISO 639~~:2023~~ provides a framework for identifying the individual languages used in an LR. The ISO 21636 series presupposes and complements ISO 639 in that it extends the language coding framework in order to allow for the identification of different types of language varieties (e.g. geographical, social, modal~~, etc.).~~). The identification of language varieties can then be included in general metadata, library metadata and archival metadata for describing LRs (which may also include technical information, time and location of recording, and similar general information, which are not included in the ISO 21636 series).

The conceptual framework developed in this document for dealing with linguistic variation respects the major approaches represented in the linguistic literature without simply reproducing them. The framework is closest though in general orientation and in a number of details, such as the role assigned to idiolects, to work of a type represented by Lieb[6] ~~(1993) .~~.

~~This document comprises:~~

~~— terms and definitions underlying a general conceptual framework to coherently deal with language-internal linguistic variation;~~

~~— terms and definitions for a set of dimensions for identifying and describing language varieties.~~

The metadata categories and values addressed in this document can be candidates for a future fine-grained coding of language varieties based on the comprehensive principles of the ISO 21636 series. Thus,

5

this document ~~complies with the "Recommendations on software and content development principles 2010", and~~ fits within the general framework of <u>the</u> ISO/IEC 11179 <u>series</u> for metadata~~.~~<u>.</u>

Stakeholders include, but are not limited to:

— information and communication technologies (ICTs) industry (including ~~language technologies~~<u>LTs</u>);

— libraries;

— the media industry (including entertainment);

— internet communities;

— people engaging in language documentation and ~~preservation~~<u>preservation</u>;

— language archivists;

— researchers (linguists, in particular sociolinguists, ethnologists, sociologists, etc.);

— people and institutions providing language training;~~ and~~

— emerging new user communities.

It is anticipated that these stakeholders will need to refer not only to a certain individual language, but also to a certain language variety, for instance for oral human-computer interaction, or for tailoring a certain LR or LT to the needs and specific environment of a target user group. An initial step towards achieving the needed specificity involves the ability to identify the dimension(s) of linguistic variation internal to individual languages involved, and the respective relevant language varieties. A conceptually sound uniform framework of reference as developed in the ISO 21636 series is superior to the proliferation of different individual ad-hoc solutions.