Content interoperability as a prerequisite for re-using and re-purposing items of structured content as learning objects in eLearning – Seen under a standardisation perspective

CHRISTIAN GALINSKI, BLANCA STELLA GIRALDO PÉREZ
Infoterm

KEYWORDS: structured content and unstructured content, verbal and non-verbal representations, appellations, morphology and morphemes, phraseology and phrasemes, collocations and multi-word terms, high complexity learning objects (HC-LO) and low complexity learning objects (LC-LO), language for general purposes (LGP) and language for special purposes (LSP), persons with disabilities (PwD) and assistive technologies, standards and standardization, certification, data models and data modelling, federated repositories, quality and interoperability requirements

1. STRUCTURED CONTENT

"Content" here is seen as ***structured content*** at the level of lexical semantics comprising linguistic and non-linguistic representations of concepts (understood in science theory as a kind of "immaterial objects"). These representations can be designative (such as designations in terminology: comprising terms, symbols and appellations) or descriptive (such as various kinds of definitions, explanations or non-verbal representations), or hybrid. So far non-verbal designations and representations of concepts as well as appellations (i.e. proper names representing individual concepts) have been under-represented in terminology theory and methodology. But they can be very important for designing learning objects (LO) in certain domains or fields of application – not to mention for language learning and translation.

Equally important for designing multilingual, multimodal and multi-purpose LO are

a) ***similarities*** of LGP and LSP entries:

- The designative representation of the entry (LGP lemma or LSP term) may be one word, a compound word or a multiword entity;
- The verbal designative representation may be supplemented by a non-verbal designative representation (e.g. gesture, mimics, Blis-

symbolics[1] etc.), or – if required – even be replaced by it (e.g. graphical or other non-verbal symbol);

- Each entry representing one concept has a unique entry identifier (which may for instance point to occurrences in text corpora);
    - In LSP entries each designative representation should have its own item identifier, which among others would make it possible to trace conceptual change over time,
    - In LGP entries each *meaning* of a designative representation should have its own item identifier, which makes multilingual entries possible,
    - In analogy to *term autonomy* in terminology, *representation autonomy* applies to LSP and LGP as well as to non-verbal representations,
    - Descriptive representations LGP (including explanations, contexts etc.) can be treated in analogy to those in LSP (explanations, contexts etc.), but exclude definitions in the strict sense,
    - There may be additional fields for notes, examples, sources etc.
- A verbal designative representation representing both a LGP as well as a LSP concept may have – depending on the degree of quasi-synonymy – to be covered by two entries, with proper cross-references.

b) ***pronunciations*** of verbal items of entities of structured content, which should be foreseen in any entry at least potentially.

c) ***non-verbal designative representations and non-verbal descriptive representations***, which should be foreseen in any entry at least potentially, because they may be – depending on the domain or field of application – equally important to verbal ones and sometimes even preferred representations.

d) ***components of entities of structured content***, such as the morphemes of verbal designative representations or elements of a definition, explanation, context etc. They should be marked / tagged in order to facilitate cross-referencing or re-using as another LO (for instance prefixes and suffixes in medical nomenclature).

e) ***larger entities of structured content***, such as idioms, LGP collocations or LSP phrasemes, as well as metaphors, which can be formally treated similar to verbal designative representations in LGP (with proper cross-references to the relevant elements of the respective entity). At this

[1] Blissymbolics is an ideographic writing system for cognitively impaired persons. Each of the several hundred basic symbols represents a concept, which can be composed together to generate new symbols that represent new concepts. Blissymbols do not correspond at all to the sounds of any spoken language.

level, the further refined differentiation of the above is necessary in corpus linguistics, but probably not for LO. It may also be of minor importance, whether the meaning is non-compositional or compositional. There are tools in corpus linguistics to identify and extract also such larger entities of structured content.

f) the provision of placeholder fields or links to the respective ***designative representations for PwD***, such as in Braille, sign language or Blissymbolics.

Structured content resources at the level of lexical semantics so far were seen as mainly comprising lexicographical data, terminological data and other kinds of concept representations, including a few non-verbal ones, such as visual symbols (e.g. public symbols). But there may be also acoustic / audible symbols, haptic / tactile symbols, and others, which, in terminology management could occur as ***designations*** or even ***concept descriptions*** (such as non-verbal representations (ISO 10241-1:2011)). There may be further information – and the respective data categories – required for a systematic approach in managing structured content at large.

In the light of the above, we can differentiate the kinds of structured content at the level of lexical semantics as follows:

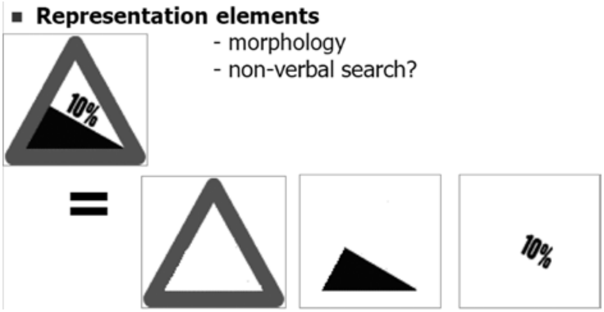
- Lexicographical data, such as:
  - word entities (including compounds etc.),
  - morphemes (morphology),
  - collocations, metaphors;
- Terminologies and similar kinds of language resources and other content resources, such as:
  - nomenclatures, taxonomies, typologies, glossaries, vocabularies etc.,
  - terminological morphemes (morphology),
  - terminological phrasemes (phraseology),
  - proper names of all sorts as used for instance as items in different kinds of directories,
  - graphical symbols and other non-verbal designations,
  - (product) properties, characteristics, attributes etc.;
- Thesauri, classification schemes (ISO/DIS 22274:2011), keywords and other kinds of documentation languages (or controlled vocabularies);
- Encyclopaedic (knowledge) entries, covering among others:
  - knowledge-enriched terminological entries,

  - (explained) proper names and other kinds of data closely related to proper names;
- Ontologies, topic maps and other kinds of knowledge-structuring systems.

Like in terminology, any of the other kinds of structured content listed above may have
- (one or more) phonetical representations, sign language representations etc.;
- graphical and other non-verbal designative representations as well as non-verbal descriptive representations (some in addition to a verbal representation, others created as non-verbal representations independent from verbal ones).

Non-verbal kinds of structured content are particularly meaningful in applications like eLearning and of vital importance in the communication:
- with and among PwD (directly or through ICT devices functioning as *assistive technologies*),
- between PwD and the devices they use, and
- among these devices.

Traffic sign designers are calling the elements of traffic signs "morphology" (in analogy to morphemes in linguistics). The same could be applied to other non-verbal designative representations. There have already been developed search methods to look for whole pictures / graphical representations by means of individual elements contained in them. The following figure shows an example for the "morphology" of graphical symbols (in this case a traffic sign):



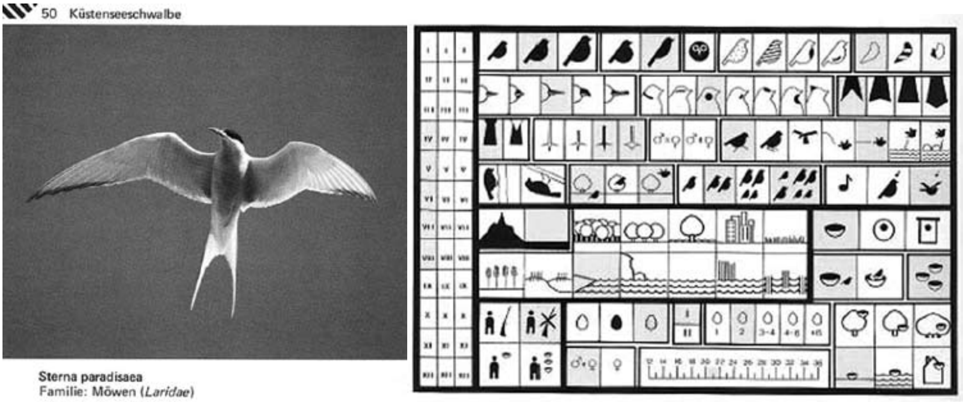
■ **Representation elements**

- morphology
- non-verbal search?

(from Schmitz' Presentation at TSTT 2006[2]; see Schmitz 2006)

[2] 3rd International Conference on Terminology, Standardization and Technology Transfer (TSTT 2006), Beijing, China, 25-26 August 2006

In certain domains (e.g. in biological nomenclatures) there are even examples for non-verbal descriptive representations replacing lengthy definitions or other concept descriptions, which could be useful also for eLearning purposes:



(from Gonnissen and Mornie 1983, bird no 50)

It only needs a comparatively short legend with explanations for the symbols (in several languages) and the graphical representation is easily understood by experts, students or hobby ornithologists alike.

All of the above kinds of structured content are becoming more and more digitally accessible today (as eContent – increasingly also through mobile devices) and may

- occur in digital texts (such as in technical documentation, scientific-technical writing etc.),
- be combined with each other or embedded in each other,
- have similar linguistic elements (letters, sounds, morphemes etc.) or different ones (such as non-verbal representations),
- form complex content items,
- need to be *integrated* or made *interoperable* with others in certain applications.

At present, however, most of the existing repositories of structured content are not consistent within a repository and sometimes even contradictory between different repositories. Mostly, they are not based on proper metadata and data modelling methods, and therefore not integrable, not reliable and full of deficiencies. This is inacceptable for instance in applications, which support PwD, particularly in our aging societies, where more and more people have – sometimes multiple – impairments. It further does not allow their efficient use for eLearning purposes – which again is disadvantaging PwD.

In contrast to structured content, a corpus can be described as ***unstructured content***, namely "a body of naturally occurring language" (McEnery, Xiao, Tono 2006), thereby distinguishing a corpus from word lists, dictionaries, databases. Similarly a speech corpus (or spoken corpus) is a repository of speech audio files (and text transcriptions). These definitions are insufficient when it comes to non-linguistic elements in texts (such as in technical documentation) and with respect to elements of non-verbal communication (such as gestures, mimics etc.) necessarily accompanying spoken data in real life.

If we look, where the items of structured content can be found, we recognize that most of them are not developed as a goal in itself, but are necessary elements of non-structured content, such as text corpora, speech corpora, etc. Therefore, the relation between structured content and corpora – especially with a view to making structured content occurring in non-structured content productive for instance for eLearning in the form of LO – should be further investigated. Today, the biggest corpus in the world is the Internet. However, in most cases it needs texts of assured quality for extracting LO – which is a challenge for classifying and quality rating of texts and tagging them appropriately.

2. STANDARDIZATION ISSUES REFERRING TO STRUCTURED CONTENT

Standards documents today do not only comprise ***technical standards*** in the traditional sense, but also standards for terminology, testing, products, processes, services, interfaces, data etc. Some are basic standards that have a wide-ranging coverage or contain general provisions for one particular field, others are methodology standards. The standardized terminology of a given field can be considered as basic standard. The methodology standards concerning the principles and methods of terminology work and terminology standardization consequently are basic for all terminology work and terminology standardization / unification in standardization at large and beyond. Furthermore, the types of standards mentioned above only refer to some common types, and they are not mutually exclusive; for instance, a particular *product standard* may also be regarded as a *testing standard* if it provides *test methods* for characteristics of the product in question. As content today is necessary for nearly everything in science and technology, content related standards can fall under any of the above-mentioned types of standards.

According to ISO/IEC Guide 2:2001 ***standardization*** is an activity for establishing, with regard to *actual or potential problems*, provisions for *common and repeated use*, aimed at the *achievement of the optimum degree of order in a given context*. In particular, the activity consists of the processes of formulating, issuing and implementing standards. Important benefits of standardization are improvement of the *suitability of products, processes and services for their intended purposes*, *prevention of barriers to trade* and *facilitation of technological cooperation*. The preparation of standards is based on ***consensus***, which is a *general agreement*, characterized by the *absence of sustained opposition to substantial issues* by any important part of the *concerned interests* and by a *process that involves seeking to take into account the views of all parties* (namely industry, research, public administration, consumers) concerned and to *reconcile any conflicting arguments*.

Standardization endeavours are governed by highly systemic approaches. In particular, methodology standards are aiming at generic solutions, which are appropriate in different applications, too. This also applies to methodology standards concerning structured content, such as standards for transcription and transliteration, data models, interchange formats, source identifiers (ISO 12615:2004) not to mention project management. It also applies to some kinds of standardized structured content itself, such as languages codes (ISO 639 series), script codes (ISO 15924:2004) and country codes (ISO 3166 series). In recent years aspects of interoperability of structured content have become a major issue in many fields. Therefore, standards concerning data quality, data administration, content management and workflows are increasingly becoming imperative.

There are several technical committees dealing with various – more or less generic – aspects of content interoperability, for instance:

- ISO/TC 37 *Terminology and other language and content resources*
- ISO/IEC-JTC 1/SC 32 *Data management and interchange* (especially its WG 2 *MetaData*);
- ISO/IEC-JTC 1/SC 36 *Information technology for learning, education and training;*

- ISO/TC 184 *Automation systems and integration* (especially its SC 4 *Industrial data*);
- ISO/TC 46 *Information and documentation.*

The standards developed by these committees do not yet take into account the specific requirements of eLearning with respect to content interoperability of entities of structured content at the level of lexical semantics.

ISO/TC 37 was the first committee to take ***multilinguality*** fully into account (in fact as one of the basic principles of all its standardization efforts, which is concept-oriented, i.e. language-independent = multilingual from the outset) – other technical committees have followed suit, but often they are not sufficiently respecting this principle in practice. Today even lexicography has taken a turn towards concept-orientation (=multilinguality), as can be seen from the products of several dictionary publishers and large-scale online dictionaries accessible on the Internet. This should also apply to learning objects (LO) at the level of lexical semantics.

Given the fact that content interoperability can only be achieved on the basis of commonly accepted rules, viz. international standards, there is a need for cooperation and coordination of the most important standardizing activities with respect to content interoperability. New standardizing activities for various aspects of content interoperability

- are driven by technical developments in the direction of web-based cooperative / participatory content creation through ICT (in the form of social networks, cooperation platforms, mobile communication etc.) on the one hand;
- will have a big impact on the various kinds of content and knowledge management (especially in eApplications, such as eLearning, eAccessibility&eInclusion, eHealth, multilingual product data management in eBusiness etc.) on the other hand.

So far the standardization efforts in ISO/TC 37 *Terminology and other language and content resources* with respect to structured content focused on methodology standards related to

- ***Data categories*** (not quite identical with metadata) used in the conceptual design of the entries of structured content;
- ***Data models*** and ***data modelling methods***;

- ***Metamodels*** to make competing data models interoperable;
- Applications of the above;

and to some extent on standardizing those kinds of content, which are of relevance to the TC.

If ontologies in the meaning of knowledge representation are included, the above-mentioned metamodels need to be extended towards ***meta-ontologies*** and even a ***meta ontology language*** (ISO/CD 17347:2012), in order to provide the possibility to make ontologies interoperable. In this connection an ontology is seen as a "formal, explicit specification of a shared conceptualisation" (Gruber, June 1993), which represents a shared vocabulary and taxonomy that models a domain – that is, the definition of concepts and other information objects, as well as their properties and relations. An ***ontology language*** – different from mere knowledge representation – provides a metamodel for such formal, explicit specifications of a shared conceptualisation.

Nearly all TCs in ISO and IEC standardize the most important terms and definitions of their respective domain or subject. Some also standardize other kinds of language and content resources.

3. VOLUMES OF STRUCTURED CONTENT: TERMINOLOGY AND OTHER LANGUAGE AND CONTENT RESOURCES

According to ELRA (European Language Resource Association) ***language resources*** are:

- text corpora,
- speech corpora,
- (lexicographical data and) terminologies.

In a recent article (Galinski, Reineke 2011), an attempt was made to quantify the volumes of lexicographical and terminological entries in an ever increasing number of domains or subjects. The lexis of LGP in the highly developed languages may comprise up to 500,000 lexemes (including a considerable share of terminology). However, the total number of scientific-technical concepts across all domains or subjects may well comprise ~100–150 million. The number of identifiable chemical substances alone has passed the 60 million mark in 2011. In the light of these figures

- Content interoperability is a prerequisite for avoiding a huge duplication of efforts;
- The ISO/TC 37 approach is becoming more and more essential;

- New approaches and methods need to be developed and existing ones adapted;
- Most of the present tools to manage terminologies are insufficient;
- The necessity for standardization – especially with respect to standards-based quality – will increase.

This is particularly important with respect to the fact that many or most terminological entries (or other items of structured content) are potential LO in eLearning.

In the above figures ***proper names and other kinds of appellations*** are not included, although in eLearning they are indispensable data (representing individual concepts). Depending on the language (or script) and the application area, they may

- Have different (similar or not similar) language versions;
- Be pronounced differently in different languages;
- Have to be transcribed into different writing systems;
- Have to be "translated" into some languages;
- Be subject to special legal conditions (such as trade marks).

The volumes of existing proper names and other kinds of appellations are uncountable – there may be several hundred million. Even here non-verbal representations may occur, such as graphical logos etc.

4. NEED FOR FEDERATED REPOSITORIES OF LEARNING OBJECTS DERIVED FROM STRUCTURED CONTENT

Counting all items of structured content at the level of lexical semantics as outlined above, which at least potentially could become LO, we may well arrive at a figure of several hundred million items. Kelly Washbourne of Kent State University once stated with regret "There is unfortunately no cure for terminology; you can only hope to manage it." This statement also applies to most kinds of structured content – not only to the linguistic ones – and especially to LO at the level of lexical semantics. Therefore, we have to extend traditional methods and tools towards new approaches, which facilitate the efficient management of the quantities and quality of structured content.

Management comprises methods, tools, processes and cooperation & coordination, as well as control mechanisms and control tools. Only if all of these are standards-based, there is a chance for enhanced interoperability of structured content developed and maintained in content repositories.

Not least for the sake of getting more ***(federated) repositories of structured content*** developed and maintained under ***quality requirements***, the situation outlined above calls for more

- methodology standards,
- content management standards,
- workflow standards (particularly with respect to web-based cooperative / participatory development of structured content),
- quality assurance standards (e.g. ISO/TS 8000-1:2011),
- database technology standards,
- standards-based verification, validation and certification schemes / tools.

Federation of repositories of structured content can have several dimensions. In a more static and passive way a repository can automatically "inform" other repositories of any modifications, which then trigger processes of dealing with these modifications. In a more dynamic way the modification of a given repository is either directly implemented in federated repositories – of course after proper validation – or the affected item in another repository draws on the modifications in the repository from where they emerged.

EXAMPLE: a display field in a product data repository reading "LENGTH 5.4 cm" is composed of the three elements:

- the metadata LENGTH for products, such as a nail or screw, having a length measured in cm;
  this metadata LENGTH has a metadata ID and a description in a metadata repository (being an authority file) – it cannot be confused with LENGTH for micro-products measured in nanometres or large products measured in metres;
- the variable value 5.4 (which, if in inch would be 2.13);
- the unit "cm", which is taken from another metadata repository (being an authority file) having its ID and a description, which may also cover conversion routines into the respective unit of the imperial system, namely "inch".

Most of the standardized structured content in ISO and IEC could theoretically become high-quality LO. Some kinds of structured content, such as basic ***terminology***, ***coding systems*** (e.g. for names of countries, currencies, languages or safety symbols), ***graphical symbols*** etc. are so important that the content items themselves are internationally standardized. In the ISO/CDB (***ISO Concept DataBase***) the following standardized content is included:

- Terminology,
- Codes,
- Graphical symbols.

It had been intended to include also:

- Quantities and units,
- Data categories (metadata),
- Product classification data,
- Product property data,
- Chemical information,
- Communication tools for certain PwD, such as Blissymbols, sign language content items (incl. sign language notation) etc.

The International Standard ISO 10241-1:2011 *Terminological entries in standards – Part 1: General requirements and examples of presentation* was developed by ISO/TC 37 with a view to include also other types of standardized structured content in the ISO/CDB. It is based on the International Standards ISO 704:2009, ISO 860:2007 and ISO 15188:2001. For terminology standardization in ISO and IEC the International Standard ISO 10241-1 is mandatory. It is not only referred to in the ISO/IEC Directives, but also applied by many terminology standardizing or harmonizing organizations in the world.

With a view to future needs of *adopting* individual standardized terminological entries, the International Standard ISO 10241-2:2012 *Terminological entries in standards – Part 2: Adoption of standardized terminological entries* has been developed. This standard will be a milestone in the direction of making standardized structured content more interoperable – also in cooperative / participatory environments.

5. LEARNING OBJECTS DERIVED FROM STRUCTURED CONTENT

According to ISO/IEC 2382-36:2008 (36.05.01) a **learning resource** is an "entity that can be referenced and used for **learning**, education and **training**". In this contribution a learning object (LO) at the level of lexical semantics is the smallest possible learning resource (LR). According to the learning object metadata (LOM) standard (IEEE 1484-12.1:2002, 3.6) a "***learning object*** [For this Standard] is defined as any entity, digital or non-digital, that may be used for learning, education or training". This contribution focuses on digital entities of structured content at the level of lexical semantics taken as LO.

The metadata (data elements) of the LOM standard, are grouped into 9 categories:

a) The *General* category is grouping the general information that describes the learning object as a whole.

b) The *Lifecycle* category is grouping the features related to the history and current state of this learning object and those who have affected this learning object during its evolution.

c) The *Meta-Metadata* category is grouping information about the metadata instance itself (rather than the learning object that the metadata instance describes).

d) The *Technical* category is grouping the technical requirements and technical characteristics of the learning object.

e) The *Educational* category is grouping the educational and pedagogic characteristics of the learning object.

f) The *Rights* category is grouping the intellectual property rights and conditions of use for the learning object.

g) The *Relation* category is grouping features that define the relationship between the learning object and other related learning objects.

h) The *Annotation* category provides comments on the educational use of the learning object and provides information on when and by whom the comments were created.

i) The *Classification* category describes this learning object in relation to a particular classification system.

Collectively, these categories form the LOM v1.0 Base Schema. The Classification category may be used to provide certain types of extensions to the LOM v1.0 Base Schema, as any classification system can be referenced.

IEEE 1484-12.1:2002 does not define how a learning technology system represents or uses a metadata instance for a learning object.

The following reflections are the outcome of a small study group on flashcards for learning Japanese vocabulary (focused on Sinojapanese characters *kanji* but covering also examples of kanji-combinations and phrases). The question was: how could they be made multilingual and used by modern devices – if possible, including mobile devices.

It was found that

- Kanji flashcards are ***high complexity learning objects*** (HC-LO) i.e. a combination of information which can be taken from several other ***low complexity learning objects*** (LC-LO);
- It needs links between LO at field level – even from parts of a given field, if necessary;
- The emerging data model must be multilingual from the outset in order to allow for cooperative / participatory development and maintenance of such LO;
- A connection to text corpora would increase the capability to extract vocabulary examples, usage examples etc. in a systematic way;
- The information of any LO is centred around a ***core element***, which can be used for browsing in texts or databases for identifying and extracting further information.

In general it was felt that a system for developing and maintaining LO of that kind could be based on the International Standards of ISO/TC 37. However, the ISO/TC 37 methodology would need some extensions, if applied to LO derived from structured content.

In any case, a kanji flashcard as shown below is a HC-LO which can be constructed of two or more LC-LO depending of the need. Whether HC-LO or LC-LO, their core elements can serve as a link to text corpora, which would allow to further add usage information, such as collocations, contexts / co-texts, metaphors, on one hand, and develop exercises, test questions, etc. from these contexts / co-texts in a systematic way on the other hand.

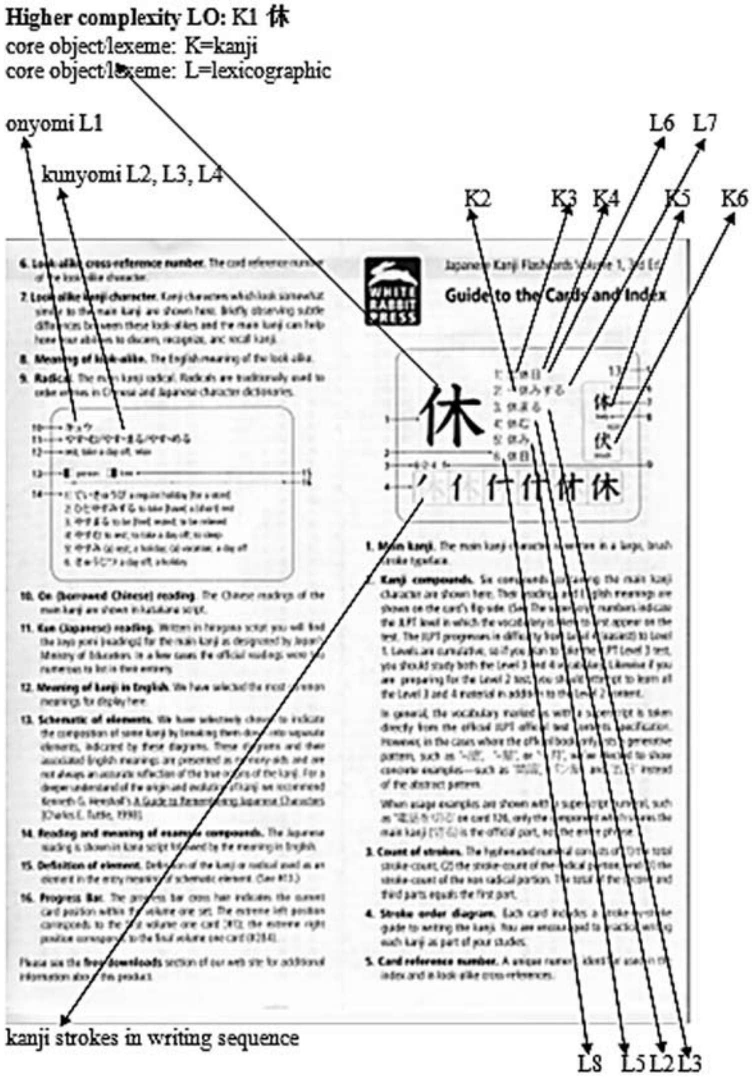
In the following figure, a flashcard for learning Sino-Japanese characters *kanji* can be decomposed into a number of HC-LO or LC-LO. Kanji flashcards have proven to tremendously increase the speed of learning kanji as well as the effectiveness of their memorization. A kanji LO by its very nature is always a HC-LO, as it may

- Have one or more "Chinese readings" *onyomi*, depending on at which period in history the kanji in question was introduced, from which region in China, and with which meaning (not to mention that the Japanese may have added different meanings in the course of time);
- Have one or more "Japanese readings" ***kunyomi***, taking the kunyomi for the Chinese character as such or as a stem adding (one or

more) endings (mostly flexion elements or particles turning the stem into an adjective or adverb etc.), the result of which sometimes can be considered as a regular derivate, however, frequently also as idiomatic;

- Be combined with other kanji to form binoms, trinoms etc.

Any onyomi or kunyomi LO (some of which semantically overlap) for a kanji LO again can be taken as a HC-LO, if they have different meanings (each at a level of lower complexity).


Higher complexity LO: K1 休
core object/lexeme: K=kanji
core object/lexeme: L=lexicographic
onyomi L1
kunyomi L2, L3, L4
L6 L7
K2
K3 K4
K5
K6
kanji strokes in writing sequence
L8 L5 L2 L3

(from Hodges and Okazaki s.a.)

In the figure above, the onyomi *kyû* (as L1 for K1 休) may carry any meaning of the kunyomi readings, but *kyû* can only be used for the kanji, if it occurs in kanji-combinations (or – which is an exception – if the kanji stands alone as an abbreviation). *Kyû* as LO would only make sense, if a learner wants to learn all kanji having the onyomi *kyû* for some reason or other. There are many kanji, where the onyomi represents one or more concepts and, therefore, each kanji+onyomi can be taken as a LO.

L2 休む *yasumu*, L3 休まる *yasumaru*, L4 休める *yasumeru,* L5 休み *yasumi* represent different – though semantically related – Japanese readings *kunyomi* of the kanji in question, which beyond their derived nature have an idiomatic meaning qualifying each of them as individual LC-LO of their own. However, if a LO of this kind has more than one meaning, any of these represents in principle a LC-LO. 休 in kunyomi readings only stands for the stem *yasu*. However, there are other stems read *yasu* with a different meaning written by different kanji, such as 安 in *yasui* 安い, which could lead the learner to other kanji LO having the Japanese reading *yasu*.

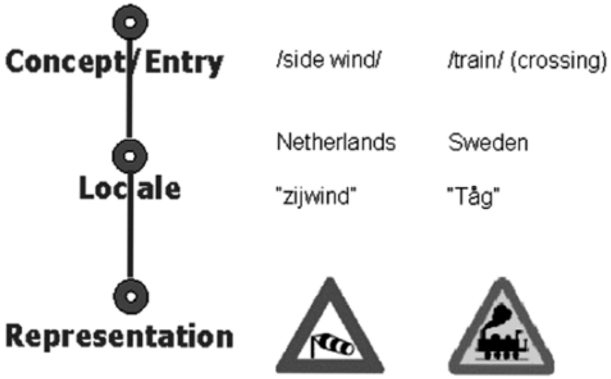
L6 定休日 *teikyûbi*, L7 一休みする *hitoyasumi suru* and L8 休日 *kyûjitsu* represent the meaning of compounds of kanji (binoms, trinoms etc. having or not having endings) with onyomi, kunyomi or mixed reading. Thus, the core element of the HC-LO K1 (休 kyû) of this kanji flashcard can

- be combined with two or more other kanji to form lexicographical LO, such as L6 *teikyûbi*, L7 *hitoyasumi suru* and L8 *kyûjitsu*. If any of them has more than one meaning, they would have to be taken as two or more LOs;
- point to other kanji HC-O, such as K2 一, K3 定, K4 日;
- be linked to look-alike kanji, such as K5 体 and K6 伏 or other kanji of historic or other relevance.

Any LO of the kinds mentioned above should have foreseen a placeholder slot for:

- pronunciation (or pronunciations, because there may be two or more; if there are homophones, pointers should lead to the respective LO);
- sign language and other kinds of AAC (alternative and augmented communication) means.

The highly complicated Japanese language and script has been chosen to illustrate the method of composing and decomposing LO at the level of lexical semantics. Taking Japanese, a complicated language with a complicated script, as a basis for designing the data models of LO for a LO repository at the level of lexical semantics has advantages. It may make it easier to deal with phenomena in the world of non-linguistic symbols, such as in the figure below.



In some countries there are even variants of the same traffic sign, depending where it is used in terms of position on the road or whether as a traditional traffic sign or in a variable message sign (VMS) (Galinski 2011).

Most of Japanese kanji occur – often with irregular readings – in appellations, i.e. proper names of people, places, organizations, buildings and other objects. Especially in languages with non-Latin writing systems the correct writing and reading of proper names is often posing great difficulties to the learner. Not to mention the correct sorting of names in such writing systems. Flying over Siberia to China one can read on the flight monitor the Chinese name for Novosibirsk starting with 新*xin*, which means 'new', translating the name element *novo* into Chinese. The rest of the name sibirsk=Siberia is transcribed by Chinese characters. Thus, proper names may have to be transcribed or (partly or fully) "translated". In any case proper names – equally to terminological and lexicographical entities – can be important LO and can be treated with the same approach outlined above.

The approach of ISO/TC 37 can serve as a model for all kinds of structured content. But as already mentioned above, the *Data Category Registry for language resources* (DCR) (ISO 12620:2009) of ISO/TC 37 should and in fact can be extended towards other kinds of structured content beyond lexicographical and terminological data, as well as towards new eApplication needs, such as for eLearning and eAccessibility / eInclusion purposes.

For those, who are frightened by the complexity, there may be a – not so novel, but in reality rarely applied – way out: federation of resources of structured content. Applying the philosophy of Dublin Core, a certain stratum of data categories (being the same for different kinds of structured content) should be standardized and implemented in any resource thus guaranteeing content interoperability. The full degree of complexity is not implemented in one super-resource, but distributed over several resources of different kinds of structured content. This way of ensuring content interoperability having the re-purposing of entities of structured content in mind still needs further investigation.

6. FUTURE STANDARDIZATION NEEDS

In the light of the above, ISO/DCR may need additional data categories. ISO/DCR today contains the basic and many extended data categories for terminological and lexicographical data. The ***terminological data model*** as well as the ***lexicographical data model*** is based on them. The use of data categories (not quite the same as *metadata* or *data elements*) seems to be highly appropriate for structured content in general and for LO at the level of lexical semantics in particular.

Besides, data categories can be distinguished into ***primary data categories*** and ***secondary data categories*** (ISO 10241-1:2011) (or even tertiary data categories). Primary data categories refer to core data, such as term, definition etc. Secondary data categories refer among others to attributes, such as preferred, admitted or deprecated (term). Tertiary data categories refer to additional information (if they are not regarded as secondary data categories), such as those referring to the elements of the source reference of terminological data (applicable also to other kind of structured content) as outlined in ISO 12615:2004 *Bibliographic references and source identifiers for terminology work.*

Primary data categories according to ISO 10241-1 adapted to LO are (compared to Table 1 in the Annex 1):

- entry number – unique for the LO database entry;
- (metadata category:) designative representation of structured content → i.e. terms, covering also synonyms, homonyms etc. to be extended towards any kind of designative representation of structured content, each of which should have its own entity ID, thus extending term autonomy towards representation autonomy;
- (metadata category:) descriptive representation of structured content → i.e. definitions, explanations, contexts etc. to be extended

towards any kind of descriptive representation of structured content including non-verbal descriptive representations;
- (metadata category:) example → if necessary to be split up into different types of example;
- (metadata category:) note → if necessary to be split up into different types of note;
- (metadata category:) source → if necessary to be split up into different types of source and their respective conditions for use.

If the entity of designative representation of structured content is verbal, the repeatability by language applies, as well as the repeatability within language in case of synonyms (each of which also should get its own entity ID). If the entity of designative representation of structured content is non-verbal, the repeatability by field of use / application applies. There may also be a kind of repeatability within the field of use / application (such as speed limit road signs depending on their location alongside the road or in the middle of a highway in some countries, or road signs of same meaning differing in various states of the USA). There are variations of sign language within the same sign language as well as of Blissymbols.

The above shows that repeatability should be according to the "community of use" (viz. *locale* in localization) rather than according to language. Of course the *community of use* can also be a language community, but the repeatability by and within locales provides more flexibility to cover all the phenomena encountered in structured content.

7. CONCLUSIONS

Over the last 10 years the limitations of ***semantic interoperability*** under a computer science perspective have become obvious. In addition to technical (i.e. hardware- and software-related) and organizational interoperability, semantic interoperability should comprise syntactic, conceptual and *pragmatic interoperability*. ***Content interoperability*** provides an even broader and generic approach with respect to the communicative representations of information and knowledge – it also takes full-fledged re-usability and re-purposability of entities of different kinds of structured content into account (Galinski, Van Isacker 2010). Re-purposability may comprise for instance the adaptation of existing terminological entries

- as learning objects (LO) in eLearning or
- for being used by persons with disabilities (PwD) for general communication and of course also for learning purposes.

International Standards for content interoperability are the prerequisite for:
- avoiding a huge duplication of efforts,
- developing methods (incl. certification) and devices to assure content quality,
- introducing content interoperability into educational and training schemes,
- enabling many eApplications to re-use and re-purpose existing structured content extensively.

Increasingly, all kinds of structured content should be re-usable and re-purposable across system platforms. In addition, more and more web-based cooperative and distributed methods for content development should be implemented in cooperation with interrelated fields under an integrative approach. In this connection the intensified use of mobile technologies will have a huge impact: eContent is becoming mContent (mobile content).

In the course of these developments, it would be worthwhile to develop ***stronger methodological and system design relations*** between ***content resource management*** and ***corpus linguistics*** in order to make better use of
- existing and future corpora with improved features e.g. for extracting individual items of structured content in a systematic way (also taking the occurrence frequencies depending on domain, level / register and application into account so that the extraction of learning objects in context would be improved);
- items of structured content in existing and future repositories for the sake of improving the processing of corpora for additional new purposes such as didactics;
- existing and emerging methods as well as tools in fields, which so far show a low degree of methodological interrelation and integration, for the sake of mutual benefits;
- systematically (and by participatory efforts) created and maintained learning objects in particular for CLIL (content and language integrated learning).

Efforts in standardization (and cooperation in standardization) should be stepped up. The proper integration of AAC requirements in the data modelling of structured content would benefit not only those in society, who need it most urgently, but ultimately everybody. This requires a stronger emphasis on pertinent assistive technologies in ICT-related education and training as early as possible.

Participants at the ICCHP 2010[3] Conference confirmed that existing training and formal studies are not sufficient – even if certified under given certification / attestation systems – with respect to the skills and qualifications necessary for becoming familiar with the issues involved in global content interoperability and particular in eAccessibility&eInclusion. The "Recommendation on software and content development principles 2010" (see Annex 2) was formulated in a special workshop at ICCHP 2010 and thereafter endorsed by several technical committees in the field of standardization, and in 2012 by the Management Group (MoU/MG) of the ITU-ISO-IEC-UN/ECE Memorandum of Understanding concerning eBusiness standardization.

REFERENCES

Galinski Ch. 2011: A sign equals thousand words. Consistency of traffic / road signs and verbal messages. – *Infrastructure and safety in a collaborative world. Road traffic safety.* Evangelos Bekiaris, Marion Wiethoff, Evangelia Gaitanidou eds., Heidelberg, Dordrecht, London, New York: Springer, 263–284.

Galinski Ch., Reineke D. 2011: Vor uns die Terminologieflut. – *eDITion* 2/2011, 8–12.

Galinski Ch., Van Isacker K. 2010: Standards-based Content Resources: A Prerequisite for Content Integration and Content Interoperability. – *Computers helping people with special needs. 12th International Conference, ICCHP 2010, Vienna, Austria, July 2010. Proceedings*, Part I., K. Miesenberger e.a. (eds.), Berlin, Heidelberg, New York: Springer, 573–579.

Gonnissen L., Mornie G. 1983: *Bestimmen und erkennen leicht gemacht. Vögel. Die wichtigsten heimischen Arten* [Birds of Europe], translated from French by Sieglinde Summerer and Gerda Kurz, Zürich/Köln: Benziger Verlag.

Gruber T. R., June 1993: A translation approach to portable ontology specifications (PDF). *Knowledge Acquisition* 5 (2), 199–220. – http://tomgruber.org/writing/ontolingua-kaj-1993.pdf.

Hodges M., Okazaki T. *Japanese kanji flashcards.* Series 2, volume 1, Tokyo: White Rabbit Press, s.a.

Kitazawa T., Windhab R., Galinski Ch. 2007: *Untersuchung von Flashcardsystemen (elektronische Lernkarteien) für CALL (computergestützten Spracherwerb) und CASPLL (computergestützten Fachsprachenerwerb) auf Anfängerniveau. Mit dem Schwerpunkt auf Japanisch-Lernen zum Erwerb einer hohen lexikalischen Sprachkompetenz – ohne das langfristige Ziel einer mehrsprachigen Flashcard-Lernplattform zu vernachlässigen.*

McEnery T., Xiao R., Tono Y. 2006: *Corpus-based Language Studies: An Advanced Resource Book*, London/ New York: Routledge.

Schmitz K.-D. 2006: Data modeling: From terminology to other multilingual structured content. – *TSTT 2006. International Conference on Terminology, Standardization and Technology Transfer. Proceedings. Beijing, China, 2006-08-25/26.* Yuli Wang, Yu Wang, Ye Tian (eds.), Beijing: Encyclopedia of China Publishing House.

[3] 12th International conference on computers helping people with special needs (ICCHP 2010), Vienna, Austria, July 2010

REFERENCED STANDARDS

ISO/IEC Guide 2:2004 *Standardization and related activities – General vocabulary.*

ISO 639 (series) *Codes for the representation of names of languages.*

ISO/IEC 2382-36:2008 *Information technology – Vocabulary* – Part 36: *Learning, education and training.*

ISO 3166 (series) *Codes for the representation of names of countries and their subdivisions.*

ISO/TS 8000-1:2011 *Data quality* – Part 1: *Overview.*

ISO 10241-1:2011 *Terminological entries in standards* – Part 1: *General requirements and examples of presentation.*

ISO 10241-2:2012 *Terminological entries in standards* – Part 2: *Adoption of standardized terminological entries.*

ISO 12615:2004 *Bibliographic references and source identifiers for terminology work.*

ISO 12620:2009 *Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources.*

ISO 15924:2004 *Information and documentation – Codes for the representation of names of scripts.*

ISO/CD 17347:2012 *Ontology Integration and Interoperability (OntoIOp)* – Part 1: *The Distributed Ontology Language (DOL).*

ISO/DIS 22274:2011 *Systems to manage terminology, knowledge and content – Concept-related aspects for developing and internationalizing classification systems.*

IEEE (Institute of Electrical and Electronics Engineers, Inc.). LOM standard 1484.12.1-2002 *Draft Standard for Learning Object Metadata.*

TURINIO SĄVEIKUMO BŪTINYBĖ NAUDOJANT STRUKTŪRIZUOTO TURINIO VIENETUS KAIP ELEKTRONINIO MOKYMOSI OBJEKTUS PAKARTOTINAI AR PAGAL KITOKIĄ PASKIRTĮ

Daugėja internetinių turinio platformų, siūlančių vartotojams vieną ar daugelį išteklių, bet vis dar trūksta teorinio ir metodologinio tokios veiklos pagrindimo, mažai atsižvelgiama į geriausias turinio sąveikumo užtikrinimo patirtis. Tokių platformų skaičius augs ir toliau, nes vis daugiau išteklių kuriama ir naudojama taikant nepakankamai veiksmingus metodus. Kad būtų užtikrinta turinio kokybė, o pirmiausiai patikimumas, reikalingas įvairių priemonių derinimas (standartai, informacinės ir komunikacinės technologijos, sertifikavimas ir kt.).

Elektroniniam mokymuisi svarbus skirtumas tarp bendrosios kalbos (angl. LGP) ir specialiosios kalbos (angl. LSP). Tam, kad būtų galima sukurti daugiakalbius, daugiamodalinius, daugelio paskirčių mokymosi objektų duomenų modelius, leksinės semantikos lygmeniu būtina tobulinti ribotas dabartinių duomenų bazių galimybes. Reikėtų daugiau dėmesio skirti:

a) duomenų bazėse pateikiamų bendrosios kalbos ir specialiosios kalbos įrašų panašumui;

b) duomenų bazių įrašų žodinių elementų tarimo nuorodoms;

c) pateikimui nežodinių žymenų ir atvaizdų, kurie, atsižvelgiant į taikymo sritį, gali būti tokie pat svarbūs kaip žodiniai žymenys ir net labiau už juos pageidautini;

d) leksinių vienetų dėmenims, pavyzdžiui, morfologiniams elementams;

e) didesniems leksiniams vienetams, pavyzdžiui, bendrosios kalbos kolokacijoms ar specialiosios kalbos frazemoms;

f) neįgaliųjų komunikacinėms reikmėms.

Straipsnyje pateikiami argumentai dėl žemiau išvardytų dalykų būtinumo:

- standartų, kuriuose būtų pateikti pasaulinio turinio sąveikumo užtikrinimo reikalavimai ir gairės, įskaitant elektroniniam mokymuisi keliamus reikalavimus;

- koordinuotos strategijos, skatinančios tokių standartų taikymą (pavyzdžiui, pasitelkiant sertifikavimą);
- priemonių, užtikrinančių standartų laikymąsi plėtojant sistemas.

Šie standartai ir priemonės galėtų padėti suvaldyti jau dabar daugiau ar mažiau chaotišką turinio plėtrą (sukeliančią didžiulį veiklos dubliavimą) ar bent jau leistų aiškiai išskirti patikimo turinio saugyklas.

Gauta 2012-03-20

Christian Galinski
International Information Centre for Terminology (Infoterm)
Gymnasiumstrasse 50, 1190 Vienna, Austria
E-mail cgalinski@infoterm.org

Blanca Stella Giraldo Pérez
International Information Centre for Terminology (Infoterm)
Gymnasiumstrasse 50, 1190 Vienna, Austria
E-mail blancaese@gmail.com

ANNEX 1

Table 1 — Overview of data categories of a standardized terminological entry in accordance with ISO 10241-1

Primary data categories[a]		**Secondary data categories**[b] (including administrative data and usage information)
Name	**Mandatory/optional; repeatable/non-repeatable**	
entry number ...	Mandatory; non-repeatable	—
term[c] (or string of five half-high dots “·····” or other slot holder sign, ...) in the order preferred term(s), admitted term(s), deprecated term(s)	Mandatory; repeatable	grammatical information in accordance with the rules of the standardizing body: e.g. gender, number, part of speech
		language code or script code, or both
		geographical use (e.g. country code)
		pronunciation
		normative status
letter symbol ...	Optional (unless the letter symbol is internationally standardized); repeatable	language code or script code, or both
		geographical use (e.g. country code)
		normative status
graphical symbol ...	Optional (unless the graphical symbol is internationally standardized); repeatable	geographical use (e.g. country code)
		normative status
definition ...	Mandatory (unless a non-verbal representation is used by convention within the respective domain or subject); non-repeatable	domain or subject, if necessary

Primary data categories[a]		**Secondary data categories**[b] (including administrative data and usage information)
Name	**Mandatory/optional; repeatable/non-repeatable**	
non-verbal representation ...	Mandatory (if exists – complementing a definition or used instead of a definition by convention within the respective domain or subject); non-repeatable	—
example ...	Optional; repeatable	—
note to entry (including note to term, letter symbol, graphical symbol, definition, context, non-verbal representation, example, a given language section of a multilingual terminological entry or the entire terminological entry) ...	Optional; repeatable	—
source of entire terminological entry (including source of term, letter symbol, graphical symbol, definition, context, non-verbal representation, example) or any language section of a multilingual terminological entry	Optional (unless the terminological entry or a language section of a multilingual terminological entry is taken from an external authoritative source); repeatable	additional information relating to the source, such as page number or clause number

[a] All primary data categories except the data category "entry number" are repeatable by language and, therefore, apply to multilingual standards. Additional rules for terminological entries in a multilingual terminology standard are given in Clause 7.

[b] All secondary data categories are optional except where they are crucial for disambiguation (...) and in cases where multilingual information is included in one terminological entry, in which case the primary data categories shall be complemented by a code for names of language in accordance with ISO 639, if necessary in combination with codes for names of countries in accordance with ISO 3166 or codes for names of scripts in accordance with ISO 15924.

[c] For simplicity, only "term" is specified in Table 1 although other verbal designations, such as any existing synonymous terms, variants, full forms, abbreviated forms, homographs, antonyms, as well as equivalent terms in other languages are included under this data category name.

ANNEX 2

RECOMMENDATION ON SOFTWARE AND CONTENT DEVELOPMENT PRINCIPLES 2010

Formulated at the ICCHP 2010 and endorsed by ISO/TC 37 and other technical committees

PURPOSE

This recommendation addresses decision makers in public as well as private frameworks, software developers, the content industry and developers of pertinent standards. Its purpose is to make aware that multilinguality, multimodality, eInclusion and eAccessibility need to be considered from the outset in software and content development, in order to avoid the need for additional or remedial engineering or redesign at the time of adaptation, which tend to be very costly and often prove to be impossible.

BACKGROUND

In software development, globalization[1], localization[2] and internationalization[3] have a particular meaning and application. In software localization they have been recognized as interdependent and of high importance from a strategic level down to the level of data modelling and content interoperability.

In 2005 the Management Group of the ITU-ISO-IEC-UN/ECE Memorandum of Understanding on eBusiness standardization adopted a statement (MoU/MG N0221), which defines as basic requirements for the development of fundamental methodology standards concerning semantic interoperability the fitness for

- multilinguality (covering also cultural diversity),
- multimodality and multimedia,

[1] **Globalization** refers to all of the business decisions and activities required to make an organization truly international in scope and outlook. G11N is the transformation of business, processes and products to support customers around the world, in whatever language, country, or culture they require.

[2] **Localization** is the process of modifying products or services to account for differences in distinct markets. Therefore, L10N is an integral part of G11N, and without it, other globalization efforts are likely to be ineffective. The interdependence of G11N and L10N has also been coined glocalization.

[3] **Internationalization** is the process of enabling a product at a technical level for localization. An internationalized product does not require remedial engineering or redesign at the time of localization. Instead, it has been designed and built from the outset to be easily adapted for a specific application after the engineering phase.

- eInclusion and eAccessibility,
- multi-channel presentations,

which have to be considered at the earliest stage of

- the software design process, and
- data modelling (including the definition of metadata),

and hereafter throughout all the iterative development cycles.

The above requirements are a prerequisite for global content integration and aggregation as well as content interoperability. Content interoperability is the capability of content to be combined with or embedded in other (types of) content items and to be extensively re-used as well as re-purposed for other kinds of eApplications. In order to achieve this capability, software must support these requirements from the outset. The same applies to the methods and tools of content management – including web content management.

RECOMMENDATION

Software should be developed and data models for content prepared in compliance with the above-mentioned requirements to facilitate the adaptation to different languages and cultures (localization) or new applications (re-purposing), the personalization for different individual preferences or needs, including those of persons with disabilities. These requirements should also be referenced in all pertinent standards.