Christian Galinski & Blanca Stella Giraldo Pérez

Infoterm[I]

THE WORK OF TODAY'S TC/37 IN THE LIGHT OF 'CONTENT INTEROPERABILITY' STANDARDIZATION AS AN ISSUE OF CORPUS PLANNING TAKING ELEARNING AS AN EXAMPLE

Abstract

This paper argues that there is a need for (1) standards to provide the requirements, rules and guidelines for global content interoperability, (2) coordinated strategies to enforce such standards (e.g. by means of certification schemes), (3) measures to guarantee that these standards are widely respected in ICT system development. "Content" here is seen as structured content at the level of lexical semantics comprising linguistic and non-linguistic representations of concepts or "objects" (incl. concepts understood as "immaterial objects"). There is a proliferation of web-based content platforms that offer users one or multiple resources on the one hand, and a lack of theoretical-methodological foundation as well as a lack of orientation at best practices and content interoperability on the other hand. A combination of means, such as standards, appropriate software, certification schemes, etc. is necessary to assure the quality – i.e. first of all reliability – of structured content. This would help to avoid a further deterioration of today's more or less chaotic development of content resources (involving a huge duplication of efforts), or at least lead to a situation where those repositories containing reliable content are clearly marked.

1 SEMANTIC INTEROPERABILITY, CONCEPT ORIENTATION AND MULTILINGUALITY & MULTIMODALITY

Structured content at the level of lexical semantics largely comprises linguistic and non-linguistic representations of concepts or "objects" (incl. concepts understood as "immaterial objects"). These representations can be designative such as designations, comprising terms, symbols and appellations, or descriptive such as various kinds of definitions, or hybrid. In this connection the non-linguistic representations of concepts, which have so far been underrepresented in terminology theory and methodology, need to be integrated into the language-independent (=multilingual) approaches for managing structured content.

Over the last 10 years, the limitations of semantic interoperability under a computer science perspective have become obvious. Further to technical (i.e. hardware- and software-related) and organizational interoperability, semantic interoperability should comprise syntactic, conceptual and pragmatic interoperability. Content interoperability provides a broad and generic approach with respect to the communicative representations of information and knowledge – it also takes into account full-fledged re-usability and re-purposability of content.

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).

This implies that not only multilinguality, but increasingly also the full range of multimodality must be covered by the data models for structured content. PwD would particularly benefit from high-quality interoperable content – whether used in eAccessibility&eInclusion applications such as ambient assisted living (AAL) or for eLearning purposes. The generic standards of the Technical Committee 37 of the International Organization for Standardization ISO/TC 37 Terminology and other language and content resources are of particular importance for content interoperability.

The Scope of ISO/TC 37 reads: *"Standardization of principles, methods and applications relating to terminology and other language and content resources in the contexts of multilingual communication and cultural diversity."* It includes the terms **content resources**, multilingual **communication** and cultural **diversity**. Content resources definitely cover content items (of structured content) which go beyond the

language means of communication. Diversity is more than just cultural diversity and also covers e.g. the needs of PwD. Under its new title *"Systems to manage terminology, knowledge and content"*, the Sub-committee ISO/TC 37/SC 3 would be an appropriate framework for dealing with standards concerning **global content interoperability**. Its scope reads: *"Standardization of specifications and modeling principles for systems to manage terminology, knowledge and content with respect to semantic interoperability."* [ISO/TC 37/AG 2011]

New standardizing activities for various aspects of content interoperability are driven, on the one hand, by technical developments in the direction of web-based cooperative/participatory content creation through information and communication technology, ICT (in the form of social networks, cooperation platforms, mobile communication, etc.). On the other hand, they 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.

There are also other technical committees dealing with various – more or less generic – aspects of content interoperability, such as:

- ISO/IEC-JTC 1/SC 32 Data management and interchange (especially its working group 2, 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 subcommittee 4, SC 4 Industrial data;
- ISO/TC 46 Information and documentation.

ISO/TC 37 was the first committee to take multilinguality fully into account (in fact as one of the basic principle of all its standardization efforts, which is concept-oriented, i.e. language-independent and thus multilingual from the outset) – other technical committees have followed suit, but often do not sufficiently respect 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 existing large-scale online dictionaries. At the user-interface, however, the data are presented as in traditional bilingual dictionaries.

2 THE 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

A corpus can be described as "a body of naturally occurring language" [McEnery, Xiao, and 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. In this sense corpora are considered as **unstructured content**. These definitions are too restricted 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.

Most of items of structured content 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 – should be further investigated.

Recently, Galinski & Reineke (2011) made an attempt to quantify the volumes of lexicographical and terminological entries in an ever increasing number of domains or subjects. The lexis of GPL (general purpose language) in the highly developed languages may comprise up to 500,000 lexemes including a considerable share of terminology. The total number of 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 important;
- new approaches and methods need to be developed and existing ones adapted;
- most of the present tools to manage terminologies are totally inadequate, and
- the necessity for standardization – especially with respect to standards-based quality – will increase.

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

- have different, but similar language versions;
- be pronounced differently in different languages;
- have to be transcribed into different writing systems;
- have to be "translated" into certain languages;
- be subject to special legal conditions (such as trade marks).

Structured content resources at the level of lexical semantics mainly comprise lexicographical data, terminological data and other kinds of concept representations.

- **Lexicographical data** cover not only words and compounds or collocations, but also morphology, idiomatic expressions etc. – and, if needed for instance in language learning/teaching (LL/LT), also phonetic transcription, pronunciation, mimics, gestures, etc.; increasingly they should also cover – if necessary – braille, sign language representations and representations in other modalities.
- **Terminological data** cover not only terms (mono-word term and multiword terms) and phraseological units, but also morphology – and if needed for instance in LL/LT also phonetic transcription, pronunciation, etc.; increasingly they should also cover – if necessary – braille, sign language representations and representations in other modalities.
- **Non-linguistic content** resources may cover visual symbols (e.g. graphical/pictorial symbols as seen in traffic signs), acoustic/audible symbols, haptic/tactile symbols, and others.

These concept representations do not only comprise the designations, but also concept descriptions, such as definitions and non-verbal representations [See: ISO 10241-1:2011], as well as other information necessary according to the data categories needed for a systematic approach in managing structured content.

Further distinguishing different kinds of structured content at the level of lexical semantics, there are:

- lexicographical data;
- terminologies and similar kinds of language resources, such as
    - nomenclatures, taxonomies, typologies, glossaries, vocabularies, etc.,
    - terminological phraseology, morphology,
    - proper names, addresses and other items of different kinds of directories,
    - graphical symbols and other non-linguistic representations,
    - (product) properties, characteristics, attributes, etc.
- thesauri, classification schemes [See: ISO/DIS 22274:2011], keywords and other kinds of documentation languages (or controlled vocabularies);
- encyclopaedic (knowledge) entries, covering among others
    - knowledge-enriched terminology entries and
    - proper names together with other kinds of data closely related to proper names;
- ontologies, topic maps and other kinds of knowledge-structuring systems.

Some of the above kinds of structured content at the level of lexical semantics can also be represented in non-linguistic/non-verbal form in addition to a verbal designation; others are created as non-linguistic/non-verbal items of structured content independent from lexicographical or terminological representations. Non-linguistic kinds of structured content are important in applications like eLearning and of vital importance in communication:

- with and among PwD (directly or through ICT devices as assistive technologies),
- between PwD and the devices they use, and

- among these devices.

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

- occur in digital texts,
- be combined with each other or embedded in each other,
- have elements (letters, sounds, morphemes etc.),
- form complex content items.

In present reality, however, most of the existing repositories of structured content are not consistent within a given repository and contradictory between different repositories. Mostly they are not based on proper metadata and data modeling 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 suffer from multiple impairments.

In order to secure the development of more **(federated) repositories of structured** content and their maintenance 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 [See among others: ISO/TS 8000-1:2011],
- database technology standards,
- standards-based verification, validation and certification schemes/tools.

3 MULTILINGUALITY = TRANSLATION? THE LANGUAGE AND CONTENT TECHNOLOGIES

The standardizing activities of ISO/TC 37 in its early years (after 1951), were dominated by domain experts and standardization experts. Between the mid-sixties and the mid-seventies of the last century, information scientists and documentation experts joined it. In the 1980s, specialized translators got interested in ISO/TC 37, but its work was considered by many of them as interference to their "terminological competence". Linguists only joined in the late 1990s, which finally resulted in the establishment of ISO/TC 37/SC 4 "Language resource management" in 2002.

In the 1970s and 1980s, the issue of **multilinguality/multilingualism** at European level was closely linked to translation and translation technology parallel to similar efforts in the USA and Japan. A study report to the Directorate General for Translation of the European Commission [Language Technology Centre 2009] states:

> In the 6th Framework Program, the European Union spent 135 million Euros on multi-modal interfaces and language technology, i.e. roughly 15 million Euros per year on language technology (Lazzari, 2006). The European Commission understood early on that language technology is an economic, political and cultural necessity in Europe. Breaking the language barrier would boost communication and the economy. (...) For the European Union, with its 23 official languages and many more spoken languages, the availability of fast, reliable and cheap translation is a necessity, and translation technology should be considered as strategically important.

The study concludes that beyond the European market, the demand for language technologies will be high in Japan, China, Korea and India.

Today the focus covers a much broader range of issues, including ICT in general, digital content (eContent), content infrastructures, etc. In particular content infrastructures are developing as part of the rapidly evolving information and communication society. Advanced computing technologies are trying to bridge the gap between computer intelligence and human intelligence, to enable computers to be more effective in areas for which they are desperately needed, such as knowledge management and text analytics. These areas of activity and research require sophisticated approaches to the encoding and

management of content resources (…) and multilingual information frameworks. [Adapted from ISO/TC 37/AG 2011]

In nearly all eApplications, the use and re-use of all kinds of structured content across different technical platforms is becoming a must. Besides, today, strongly heterogeneous content is still more the rule than the exception. Whereas in the past the development focus was on tools (i.e. devices, computer hardware and software), it is increasingly recognized now today that communication and content are what really counts. That is why work on content related aspects is increasing all across standardization.

In any case the research focus since the late 1990s has shifted towards language technologies at large and progressively also to content (and the related technologies) in all its guises, the use and re-use of content in various eApplications, the development of new applications (incl. eLearning, eHealth, eAccessibility&eInclusion), new forms of web-based cooperative / participatory methods for content creation, maintenance and use etc. These developments have an impact on standardization activities in ISO/TC 37, which are subject to a number of push and pull factorsII:

- ICTs are developing in the direction of mobile technology (inevitably enhancing also the development of cloud computing and crowd sourcing) and social web approaches (where cooperative approaches prevail);
- new applications emerge, while others are converging;
- higher demands for content and service quality are made by the user;
- new markets for content-related data and services are emerging, etc.

Needless to say, the above-mentioned developments have a big impact on software development, as can be gathered from the documents MoU/MG/05 N0221 (2005) and MoU/MG/05 N0222 (2006). Modularity and comprehensive interoperability, capability for multilinguality and multimodality based on open standards are increasingly required. In this connection, the "Recommendation for software and content development principles 2010" (see Annex) launched by Infoterm at the 12th International Conference on Computers Helping People with Special Needs (ICCHP 2010) and endorsed among others by ISO/TC 37, will hopefully have its intended impact.

4 STANDARDIZATION AND CERTIFICATION

While in the past, development focus was on tools, it is increasingly recognized today that communication is ultimately the most important issue. Therefore, metadata, data models, messages, protocols, conversions of all sorts, multilinguality, multimodality, cultural and other kinds of diversity, design for all (DfA [see ICTSB 2000]), etc., have become the objective of standardization efforts related to structured content in industry (first of all in eBusiness), by specialized organizations or in public institutions. No wonder that the number of standards for content-related services and structured content is growing exponentially. Learners in education&training and persons with disabilities (PwD) are among those who will benefit the most from the development of content related standards taking their needs into account.

4.1 OVERVIEW OF CONTENT RELATED STANDARDS

Technically speaking, content can be subdivided into structured content (such as content recorded in database structure) and unstructured content (such as running text or streaming information). From a technical point of view, Content in content management approaches comprises:

- text (textual data),
- graph (graphical/visual data),
- sound (audible/audio data),
- multimedia (including video).

This subdivision proved insufficient from an inter-human communication perspective – not to mention the increasing need to include multilinguality and multimodality in databases for various purposes.

So far, the standardization efforts in ISO/TC 37 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 modeling methods**;
> **meta models** to make competing data models interoperable;
> applications of the above;

and to some extent on standardizing content itself, which is of relevance to the Technical Committee.

If ontologies in the sense of knowledge representation tools are included, the above-mentioned meta models need to be extended towards meta-ontologies and even a meta ontology language [See: ISO/CD 17347:2012] in order to provide the possibility of making ontologies interoperable. In this connection, an ontology is seen as a "formal, explicit specification of a shared conceptualisation" [Gruber, 1993], which represents a shared vocabulary and taxonomy that models a domain — i.e., 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 meta model for such formal, explicit specifications of shared conceptualisations.

4.2 STANDARDIZED CONTENT

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. Nearly all TCs in ISO standardize the most important concepts of their respective domain or subject. Some others standardize also other kinds of language and content resources. ISO/TC 37's standardized structured content is largely contained in ISO 1087 *Terminology work – Vocabulary*, multipart ISO 639 Codes for the representation of names of languages and, if one includes also the data categories, in the Data Category Registry (ISO/DCR).

In the **ISO Concept DataBase** (ISO/CDB), the following standardized content is included:

- terminology,
- codes (e.g. for the names of countries, currencies, languages),
- graphical symbols.

It had been intended to include also:

- quantities and units,
- data categories,
- product classification data,
- product property data,
- chemical information,
- communication tools for certain PwD, such as BLISS symbols, sign language content items (incl. sign language notation), etc.

With a view to present and future types of structured content in the ISO/CDB the International Standard ISO 10241-1:2011 Terminological entries in standards – Part 1: General requirements and examples of presentation has been developed by ISO/TC 37. It is based on the International Standards ISO 704:2009, ISO 860:2007 and ISO 15188:2001. ISO 10241-1 is mandatory for terminology standardization in ISO and the International Electrotechnical Commission (IEC). 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 Terminological entries in standards – Part 2: Adoption of standardized terminological entries is under development. This standard will be a milestone in the direction of making standardized structured content more interoperable.

4.3 STANDARDIZED DATA CATEGORIES OR METADATA

ISO 12620:2009 *Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources* is the result of a long discussion in ISO/TC 37 about the basic data categories for terminological data and the **terminological data model** to be 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.

Data categories in this connection can be sub-divided into **primary data categories** and **secondary data categories** (According to 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 sources of terminological data (applicable also to other kinds of structured content) as outlined in ISO 12615:2004 Bibliographic references and source identifiers for terminology work.

The data category approach in ISO/TC 37 may serve as a model for all kinds of structured content. In future, the ISO/DCR might and should be extended towards other kinds of structured content beyond lexicographical and terminological data, as well as towards new eApplication needs, such as in eLearning and for eAccessibility&eInclusion purposes.

4.4 STANDARDIZED DATA MODELS AND DATA MODELING METHODS

The ISO/TC 37 **data models** and **data modeling methods** for terminologies are based in general on the International Standards ISO 704:2009 Terminology work – Principles and methods, ISO 860:2007 Terminology work – Harmonization of concepts and terms and ISO 15188:2001 Project management guidelines for terminology standardization, and more specifically on ISO 26162:2010 Systems to manage terminology, knowledge and content – Design, implementation and maintenance of terminology management systems. The latter specifies criteria for designing, implementing and maintaining terminology management systems (TMS), including guidelines for selecting and using data categories for managing terminology in various environments. Standards of other technical committees, such as the multipart ISO/IEC 11179 are duly taken into account. For a couple of years, a group within ISO/TC 37 has also been dealing with the application of Unified Modeling Language (UMLII) in the field of terminology. The result is the Technical Report, ISO/TR 24156:2008 Guidelines for using UML notation in terminology work.

As a matter of fact virtually all kinds of structured content may

- be combined with each other or embedded in each other,
- have elements (letters, sounds, morphemes, etc.),
- form complex content items (such as composite learning objects vs. primitives).
- This is obvious in applications, such as technical documentation, translation systems, authoring tools, etc. In the future, therefore, these systems should
- be developed according to the requirements multilinguality&multimodality including those of PwD
- be developed in the form of cooperative/participatory work, and the resulting repositories should be federated.

This implies new needs for standardized data categories, as well as for data model(s) and data modeling methods to suit them. In the conceptual data model provided to ISO/CS in preparation of the development of the ISO/CDB, a future federation of similar databases was recommended. The federation could refer to different language versions as well as different domains or applications – based on the same data model.

4.5 STANDARDIZED META MODELS AND A META ONTOLOGY LANGUAGE

Among the main standards concerning meta-models in ISO/TC 37 are:

- ISO 16642:2003 Computer applications in terminology – Terminological markup framework
- ISO 24613:2008 Language resource management – Lexical markup framework (LMF)

Work continues on a meta model for multilingual terminological data and lexicographical data.

In order to fill the current gaps in modular **ontology design** and to augment current standardization efforts by an essential layer of standardised modularity and structuring guidelines, ISO/TC 37/SC 3 Systems to manage terminology, knowledge and content has adopted a new working item (ISO/NWI 17347) on Ontology Integration and Interoperability (OntoIOp) which has been conceived in the framework of a large-scale European R&D project in its work programme. The project team brings together results of the international state-of-the-art in ontology-based interoperability. This includes results from several large-scale initiatives. Thus, the proposed future International Standard ISO 17347 OntoIOp aims at bridging the above-mentioned gaps in standards and guidelines.

Partly as a result of the adoption of a description logic basis, which is typical within Semantic Web oriented information modeling, the development of more powerful and generic approaches to supporting modularity has so far been delayed. Existing meta model specifications and ontology definition standards assume that the ontologies produced are essentially compatible down to the exchange of terms and filling in of respective knowledge gaps. But this 'assumption' of ontological compatibility frequently fails to hold true, and does not match current practice nor expectations when standardization is considered across technical communities.

4.6 STANDARDS CONCERNING DATA ADMINISTRATION, CONTENT MANAGEMENT AND WORKFLOWS

Increasingly web-based, distributed and cooperative or participatory methods are applied today to manage the above-mentioned different types of structured content in such a way that the best suited "stakeholder" for a given kind of content can ensure the maintenance, updating and versioning as well as quality assurance of the content in the most efficient way possible. However, a number of additional efforts need to be made with respect to global content interoperability, such as:

- more and more high-quality **metadata** (or data categories, as they are more precisely called in the field of terminology) and the respective **data category registries**;
- **identification systems** for individual pieces of information;
- **business models** (including also solutions to certain intellectual property rights, IPR, issues).

In this connection the question of how to ensure the sustainability of repositories of structured content e.g. by means of commercial or non-commercial business models definitely needs more attention in future.

ISO/IEC-JTC 1/SC 32/WG 2 MetaData is the Working Group that formulates the methodology standards on metadata and metadata repositories, MDR. [See: multipart ISO/IEC 11179] International coordination groups recommend harmonization of a multitude of ID-systems for content items, which are a great barrier to the efficient exchange of structured content. For such exchange a systematic approach is necessary, including checking and quality assessment procedures, like verification, validation, and certification, etc. In product data management for eBusiness, among others, ISO 29002-5:2009 is proposed as an open, non-proprietary standard acting as the basis for applying an ID-system for identifying pieces of information at the most detailed level of content granularity.

This efficient use, re-use and re-purposing of structured content can only take place if two fundamental principles of content management methodology are fully realized: single sourcing and resource-sharing.

This cannot be realized through traditional approaches of committee work or individuals working for the rest of the world, but probably through new social web approaches via platforms which allow for participatory development and maintenance of structured content. However, standards for appropriate workflow design and organization are still at a stage of infancy.

4.7 STANDARDS RELATED CONCEPTS AND CERTIFICATION

With respect to content interoperability, which is a global issue, ultimately only the international standards of ISO, IEC and the International Telecommunication Union (ITU) as well as the guidelines of the World Wide Web Consortium (W3C) can guarantee the most efficient use, re-use and re-purposing of structured content across language boundaries and system platforms. When it comes to content interoperability, more consistent, coherent, and well-coordinated international standards are required rather than competing industry standards. Not least, as a result of R&D projects of the European Union (EU), the awareness of these new requirements in respect of structured content is increasing; and new ISO standards, such as ISO/CD 17347 OntoIOp, represent developments in the right direction.

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. [ISO/IEC Guide 2:2004] Standardization endeavours are governed by highly systemic approaches. In particular, methodology standards are aiming at generic solutions, which are applicable also to structured content to be used in different eApplications.

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. [ISO/IEC Guide 2:2004] Therefore, standards published by formal standards bodies are called open standards in contrast to industry standards, which are usually proprietary. Great efforts are undertaken to harmonize existing open standards at national, regional and international levels so that they do not contradict each other. In Europe the regional formal standards bodies CEN, CENELEC and ETSI are all involved in eAccessibility&eInclusion standardization in some way or other – well-coordinated with the international standardizing organizations. Naturally, this does not apply to the same degree to industry standards, although cooperation with formal standards bodies has increased during the recent years.

Certification is defined as a procedure by which a third party gives written assurance that a product, process or service conforms to specified requirements. [ISO 14050:2009] Certification involves a number of documented processes, at the end of which a documented assessment result is obtained.

The idea that data structures and data should be standards-compliant and that they may, if they are potentially highly content interoperable, be certified is relatively new. This standards compliance needs to be assessed according to validation or verification criteria, defined as policy, procedure or requirement, and used as a reference to which evidence is compared. In this connection, two closely related systematic, independent and documented processes are of relevance:

- **Verification**, for the sake of the evaluation of assertions against agreed verification criteria, applies objective evidence that specified requirements which define an intended use or application have been met. Whenever specified requirements have been met, a verified status is achieved.
- **Validation** uses objective evidence to confirm that specified requirements which define an intended use or application have been met. Whenever all requirements have been met, a validated status is achieved.

Today, attestation is often used as a collective term for the above or for the assessment of the qualification and skills of persons involved.

There are many ways to verify that requirements have been met, such as testing, performing demonstrations, carrying out alternative calculations, comparing a new design specification with a proven

design specification, or inspecting documents before one issues them. The process of validation can be carried out under realistic use conditions or within a simulated use environment.iv

The above-mentioned concepts emerged during many years of discussion of standards-based quality management systems and the requirements associated with them. In this connection, quality has been defined as the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs. The quality of data and data-related services and tools has only recently become an issue in quality assessment and certification approaches. Needless to say, the potential for quality of data and related services and tools is higher than if they are not standards-based.

Quality of something as defined by international standards is a relative concept which:

- can be determined by comparing a set of inherent characteristics with a set of requirements; *If those inherent characteristics meet all requirements, high or excellent quality is achieved. If those characteristics do not meet all requirements, a low or poor level of quality is achieved.*
- therefore, is a question of degree.

As a result, the central quality question is: How well does this set of inherent characteristics comply with this set of requirements? In short, the quality of something depends on a set of inherent characteristics and a set of requirements and how well the former complies with the latter. By linking quality to requirements, ISO 9000:2005 argues that the quality of something cannot be established in a vacuum. Quality is always relative to a set of requirements.iv

The assessment of quality management systems and services takes place by means of an audit, which is a systematic, independent and documented process for obtaining audit evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled. **Audit criteria** are a set of policies, procedures or requirements used as a reference against which audit evidence is compared. A distinction is made between **internal audits**, sometimes called first-party assessment, and **external audits**, including those generally termed second-and third-party assessment. [ISO 19011:2002] Third-party audits are conducted by external, independent auditing organizations, such as those providing registration or certification of conformity with the requirements of ISO 9001:2000 or ISO 14001:2004.

While quality assessment and certification through audits is relatively well developed in industry and services, the verification and validation of structured content and related services and tools is still in its infancy. Data on factual and physical properties can comparatively easily be validated. Other properties, e.g. those relating to "soft" criteria, such as functions, can probably only be validated if validation is performed against a limited number of pre-defined values.

Ultimately, high quality of content itself, of related tools, and of services and human competences can only be achieved if it can be assessed on the basis of standards which have been devised with certification in mind.

4.8 SPECIAL: SKILLS CERTIFICATION IN THE FIELD OF ICT

eCertification in Europe can be considered the set of processes by which an individual gains a credential in a particular ICT skill or more generally in a range of skills. Such credentials are usually granted by recognized bodies, themselves often, but not always, accredited by some governmental or official organization. In order to achieve such qualification, that individual must achieve a declared standard, judged by a formal assessment process. The whole scheme is governed by quality assurance processes, covering both the development and maintenance of the skills standard and the assessment procedures [See: CEN Workshop Agreement, CWA 16052:2009].

CWA 16052 refers to the following definitions of eCertification:
(1) "Certification often means the awarding of a certificate, or other testimonial, that formally recognizes and records success in the assessment of Knowledge, Skills and/or Competencies, as the final step in the completion of a Qualification. However, it is also used, in particular in relation to ICT Practitioner occupations, to mean the Qualification as a whole. It is important to be aware of these two meanings of Certification." [Dixon and Beier in CWA 15515:2006]

(2) "Certification is the process of formally validating knowledge, know-how and/or skills and competencies acquired by an individual, following a standard assessment procedure. Certificates or diplomas are issued by accredited awarding bodies." [Tissot: 2004 cit. in CWA 16052]
(3) "In general, ICT professional certifications are seen as a credential – the result of an objective assessment procedure run by an approved third party, in which an individual meets the performance specifications delineated in job profiles which are recognised by industry stakeholders." [CEPIS: 2007; Cedefop: 2006]

Three ISO standards are related to eCertification, viz.:

- ISO/IEC 17024:2003 *Conformity Assessment – General requirements for bodies operating certification of persons;*
- ISO/IEC TR 19759:2005 *Software Engineering – Guide to the Software Engineering Body of Knowledge (SWEBOK);*
- ISO/IEC 24773:2008 *Certification of software engineering professionals – Comparison framework.*

Whereas it has been recognized that certification provides value in both the labour and product segments of the ICT market, the HARMONISE report [CEPIS: 2007] describes over 600 often overlapping qualifications from over 60 providers as a "certification jungle", causing confusion to prospective users. The rapid growth in these industry qualifications has been driven by the market over recent years; indeed this market barely existed 15 years ago. They usually relate to a more specific set of skills, including those relating to specific products, and are generally more practical in their approach than traditional academic qualifications.

As these market certifications compete and co-exist with those of the traditional university based education system, resistance and even hostility towards these certifications exists, e.g. in academic quarters in some countries. They are seen as developing skills, not as competence based on proper education, i.e. they are considered little more than marketing aids to the commercial interests of the vendors. [CWA 16052:2009]

Participants at the above-mentioned ICCHP 2010 confirmed that available training and formal studies are not sufficient – even if certified under given certification systems – with respect to the skills and qualifications needed in order to become familiar with the issues involved in global content interoperability and particularly in eAccessibility&eInclusion.

5 NEW DEVELOPMENTS AND OUTLOOK

Content interoperability – exceeding the concept of semantic interoperability in computer science – is the capability of content items to be:

- integrated into or combined with other types of content items;
- extensively re-used for other purposes, including sub-items to be re-purposable;
- searchable, retrievable, and re-combinable from different points-of-view [Galinski & Van Isacker 2010].

This applies to all kinds of structured content, whether linguistic or non-linguistic. Increasingly, all kinds of structured content have to be re-usable and re-purposable across system platforms. In addition, progressively web-based cooperative and distributed methods for content development should be developed in cooperation with interrelated fields under an integrative approach.

International Standards for content interoperability are a prerequisite for:

- avoiding huge duplication of efforts,
- developing methods, including 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.

In the course of these developments, **stronger methodological and system design relations** between **content resource management** and **corpus linguistics** should be developed, in order to make better use of

- existing and future corpora with improved features for this purpose, e.g. for extracting individual items of structured content in a systematic way, e.g. also taking into account the frequencies of occurrence depending on domain, register and application in order to improve extraction of learning objects in context;
- items of structured content in existing and future repositories for the sake of improving the processing of corpora for present and new purposes;
- existing and emerging methods as well as tools in fields which have so far shown a low degree of methodological interrelation and integration, for the sake of mutual benefits;
- learning objects systematically created and maintained by participatory efforts, in particular for content and language integrated learning (CLIL).

With a view to the sustainability of content repositories, business models, whether commercial or non-commercial, deserve more attention than they have received so far. Efforts in standardization (and cooperation in standardization) should be increased. On the basis of pertinent standards, a whole range of certification / attestation systems may evolve. In addition, efforts need to be undertaken to introduce these developments into ICT-related education and training as soon as possible.

Annex: Recommendation formulated at the 12th International Conference on Computers Helping People with Special Needs (ICCHP 2010) and adopted by ISO/TC 37 and other committees.

RECOMMENDATION ON SOFTWARE AND CONTENT DEVELOPMENT PRINCIPLES 2010

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, globalization1, localization2 and internationalization3 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,
- 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.

[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.*

REFERENCES (scientific literature and documents)

CEDEFOP ed. (2006). ICT Skills Certification in Europe. Cedefop Dossier series 13.Luxembourg: Office for Official Publications of the European Communities

CEPIS ed.. (September 2007). Survey of Certification Schemes for ICT Professionals across Europe towards Harmonisation (HARMONISE). Project of CEPIS Council of European Professional Informatics Societies. Final report. http://www.cepis-harmonise.org

GALINSKI, Ch., VAN ISACKER, K. (2010). Standards-based Content Resources: A Prerequisite for Content Integration and Content Interoperability. In: K. Miesenberger et al eds.: Computers helping people with special needs. 12th International Conference, ICCHP 2010, Vienna, Austria, July 2010. Proceedings, Part I. Berlin Heidelberg; New York: Springer, 2010. ISBN 3-642-14096-3. pp. 573–579

GALINSKI, Ch. und REINEKE, D. (2011). Vor uns die Terminologieflut. In: eDITion 2/2011, pp. 8-12

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.

MCENERY, T. XIAO, R. & TONO, Y. (2006). Corpus-based Language Studies: An Advanced Resource Book. London/New York: Routledge.

CWA 16052:2009 ICT Certification in Europe. CEN Workshop ICT-Skills (IT profiles and curricula). CEN, Brussels

CWA 15515:2006 European ICT Skills Meta-Framework – State-of-the-art review, clarification of the realities, and recommendations for next steps. CEN Workshop ICT-Skills (IT profiles and curricula). CEN, Brussels

ICTSB (2000). ICTSB Project Team Final Report (ed.). Design for All. http://www.ictsb.org/Activities/Design_for_All/Documents/ICTSB%20Main%20Report%20.pdf

ISO/TC37/AG N 234 (2012). ISO/TC 37 Business Plan 2011

The Language Technology Centre Ltd. (2009). Study report to the Directorate General for Translation of the European Commission.

MoU/MG N0221:2004 Semantic Interoperability and the need for a coherent policy for a framework of distributed, possibly federated repositories for all kinds of content items on a world-wide scale (adopted in 2005) http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/MoU-MG/Moumg221.pdf

MoU/MG N0222:2004 Statement on eBusiness Standards and Cultural Diversity (adopted in 2006) http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/MoU-MG/Moumg222.pdf

REFERENCES (standards)

ISO/IEC Guide 2:2004 Standardization and related activities -- General vocabulary

ISO/IEC 11179 (series) Information technology – Metadata registries (MDR) ISO 10241-1:2011 Terminological entries in standards – Part 1: General requirements and examples of presentation

ISO/IEC 17024:2003 Conformity Assessment – General requirements for bodies operating certification of persons

ISO/IEC TR 19759:2005 Software Engineering – Guide to the Software Engineering Body of Knowledge (SWEBOK)

ISO/IEC 24773:2008 Certification of software engineering professionals – Comparison framework.

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

ISO 9000:2005 Quality management systems – Fundamentals and vocabulary

ISO 9001:2000 Quality management systems – Requirements

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

ISO 14001:2004 Environmental management systems – Requirements with guidance for use ISO 29002-5:2009. Industrial automation systems and integration – Exchange of characteristic data – Part 5: Identification scheme.

ISO 14050:2009 Environmental management – Vocabulary

ISO 19011:2002 Guidelines for quality and/or environmental management systems auditing

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

ISO 29002-5:2009. Industrial automation systems and integration – Exchange of characteristic data – Part 5: Identification scheme.

END NOTES

[I] Birthe Toft was the respondent to this paper at the IITF Symposium. Her comments have been taken into account by the authors as much as possible.

[II] eAccessibility concerns the design of Information and Communication Technology (explained once already)ICT) products and services so that they can be used by PwD. eInclusion aims to achieve that "no one is left behind" in enjoying the benefits of ICT. Both include the special needs of the elderly.

[III] Push factors here referring to developments within the field of terminology science and its applications, while pull factors are coming from outside, for instance out of software engineering development and eApplications in need of global content integration and interoperability.

[iv] http://www.uml.org/

[v] European Committee for Standardization (CEN), European Committee for Electrotechnical Standardization (CENELEC), European Telecommunications Standards Institute (ETSI)

[vi] http://www.praxiom.com/iso-definition.htm#Validation

[vii] http://www.praxiom.com/iso-definition.htm#Quality