Languages for Special Purposes in a Multilingual, Transcultural World

Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria

http://lsp2013.univie.ac.at/proceedings

Typology of structured content in eApplications: Under a content interoperability, quality and standardization perspective

Dr. Christian Galinski; Blanca Stella Giraldo Pérez

Cite as: Galinski, C. & Giraldo Pérez S. B. (2014). Typology of structured content in eApplications: Under a content interoperability, quality and standardization perspective. In G. Budin & V. Lušicky (eds.), *Languages for Special Purposes in a Multilingual, Transcultural World, Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria*. Vienna: University of Vienna, 405-417.

Publication date: July 2014

ISBN: **978-3-200-03674-1**

Typology of structured content in eApplications: Under a content interoperability, quality and standardization perspective

Dr. Christian Galinski
Infoterm
Austria

Blanca Stella Giraldo Pérez
Centre for Translation Studies, University of Vienna
Austria

Correspondence to: cgalinski@infoterm.org; sgiraldo@infoterm.org

Abstract. There is a proliferation of content – not only in the Internet, but also at organization level, where it is called 'big data'. Therefore, enterprises struggle to integrate content resources or at least make them interoperable. In line with this process, content has to be analyzed and low-quality content to be identified in order to be deleted or improved. This refers also to a large degree to structured content – here content entities at the level of lexical semantics comprising linguistic and non-linguistic representations of concepts. Resources of structured content were seen as mainly comprising terminological data, lexicographical data and other kinds of concept representations, including a few non-verbal ones, such as visual symbols (e.g. public symbols).

But increasingly 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). This occurs first of all in the eApplications, such as eLearning, eBusiness, eHealth etc. More and more also non-linguistic entities of structured content are subject to the multilingual requirements of industry and the need for different modalities (e.g. graphical representations made 'readable' by blind people).

The investigation reveals various kinds of structured content at the level of lexical semantics which largely coincide with the characteristics of microcontent. From the semantic point of view and generic approach of terminology, microcontent entities could be data modeled based on one open-ended data model with a core structure (based on core data categories) covering more or less all kinds of the information objects called microcontent.

Keywords. Content interoperability, eLearning, eContent, structured content, unstructured content, microcontent, data modeling, metadata, standardization, terminology.

1. Need to clarify *eContent*

Content has acquired new meanings and dimensions under the conception of *eContent*. Originally derived from *electronic content eContent* is defined as *digital content* that can be transmitted over a computer network such as the Internet. (The Computer Language Company Inc. 1981-2013) This definition implies that eContent is developed in a computer-assisted way (which does not exclude conventional output) and that it must conform to a minimum of standards, in order to be transmissible over the Web. As the development of eContent is not a goal in itself (maybe except when applied in the fine arts), the purpose of the development, e.g. for an intended application, has to be added to the definition. Besides, in the course of increasingly using eContent as a commodity, commercial and legal aspects (*inter alia* digital rights) are becoming more and more important for the distribution and use of eContent. Pertinent technical standards as well as legal norms have been developed at international, regional and national levels – not to mention new methodology standards for accommodating new commercial and legal requirements in the data models for eContent. As nearly any traditional content can be digitized with technical means and turned into eContent, 'content' today stands for traditional content as well as modern digital content and is used as such in this contribution.

Content applications, also called eApplications, are for instance eLearning, eBusiness, eHealth etc. Some of these eApplications came to use different terms for eContent, such as *intelligent content* in eBusiness or *learning material* in eLearning. While the number of different eApplications is increasing, the total amounts of content accessible on the Web are growing exponentially. Not least in order to save costs involved in content development and maintenance as well as the growing need for re-using and re-purposing available content, content increasingly has to be integrated, combined, further developed etc.. Therefore, standards are developed to ensure the integratability and interoperability of content – if possible – already at the time of its development.

In any case *content* is related to, but somehow also different from the traditional *wisdom chain* of information science consisting of: data → information → knowledge → wisdom. Content management from a theoretic-methodological point of view usually does not clearly distinguish between:

- Content and data – information – knowledge,
- Large and small objects/entities/items/units[1] of content (and the respective content resources),
- Structured and unstructured content.

This proves to be a big barrier against the integratability and interoperability of content increasingly required in more or less all eApplications.

As content is largely processed as 'digital objects' (or *digital materials* referring to any item that is available digitally) this contribution attempts to analyse content and related concepts ultimately aiming at establishing a typology of *structured content at the level of lexical semantics* in eLearning.

2. Content and related concepts

The *type of content* and where to take it from is one of the key issues that must be considered when dealing for instance with a project on a "*new methodology to design and develop learning objects (LO) based on terminological and corpus linguistic methods.*" However, the vast amount of information that can be found dealing with *content* and the diversity of areas that use the term indistinctly for more or less different concepts calls for some explanation before focusing on content entities.

More often than not the terms *data*, *information* and *content* are used interchangeably. Therefore, the relation between data – information – content will be tackled first.

Information processing theory argues that the physical world is made of information itself. Under this definition, *data* is either made up of or synonymous with physical information. Data as an abstract concept can be viewed as the lowest level of abstraction from which information and then knowledge are derived[2]. Generally speaking, information and data have much in common and are often used as synonyms in pertinent literature.

In ISO standards *data* is defined differently from the point of view of several technical committees:

- ISO 22005:2007 (3.1): recorded information;
- ISO 15784-3:2008 (3.7): information before it is interpreted;
- IEC60050-701,721:1992; ISO 16091:2002 (3.1.4): information represented in a manner suitable for automatic processing.

As for *information*, ISO standards again provide several definitions – however, without really providing a clear distinction:

- ISO/TS 25237:2008 (3.27): data set within a context of meaning;

- ISO 15531-43:2006 (3.1.15): facts, concepts, or instructions;
- ISO 22320:2011 (3.9): data that are processed, organized and correlated to produce meaning.

From the user's perspective, information is all content, while from the computer programmer's perspective, it is all data. (Boiko 2004)

In the course of the development of the eApplications, *content* has become one of those fuzzy 'terms' that needs careful scrutiny in order not to add to the confusion. In international standards, content is among others defined as follows:

- ISO/IEC 15938-5:2003 (3.3.2.9): a representation of the information contained in or related to multimedia data in a formalized manner suitable for interpretation by human means. Content refers to the data and the metadata;
- ISO/IEC 24800-3:2010 (3.1.5): data and the associated metadata;
- ISO 24531:2013 (4.11): <XML> all data between the start tag and end tag of an element.

This shows that content and metadata are closely related to each other.

ISO 9241-151:2008 defines *content* (in the meaning of web content referring to the web user interface) as "set of *content objects*" (item 3.4) and *content object* as "interactive or non-interactive object containing information represented by text, image, video, sound or other types of media" (item 3.5). Thus from a technical point of view content management takes content as *content objects* (i.e. content entities) in the form of:

- Text (i.e. textual data, incl. all kinds of alpha-numeric data),
- Sound (audio data),
- Image (graphical data),
- Video (incl. multimedia data).

Other types of media indicate that other modalities (defined in ISO 5492:2008, item 2.11, as "sensations mediated by any of the sensory systems, for example auditory, taste, olfaction, touch, somesthesis or visual modality") are not excluded. On the one hand, text, image and video (without sound) refer to the visual modality; on the other hand, different media today may occur in texts or documents, e.g. in electronic books. This reveals that *content* is not satisfactorily defined from the technical point of view. From the standpoint of semantics, the above definition is certainly insufficient.

Currently, *content management* has largely gained control of the term *content*. ISO/IEC/IEEE 26511:2011 (item 4.4, from a user documentation point of view) defines *content management* as "control of units of information with their metadata, to allow selective reuse in documents or information items with variable structures and formats". To this is added the "EXAMPLE: Content management for user documentation means management of help topics, explanations of concepts, troubleshooting procedures, compliance statements, and variables such as the names and host platforms of software products, with metadata tags that are applied to format output". It seems as if *information units* here resembles *content entities*, while "reuse in documents or information items" indicates that there are objects/ entities/items/units of higher complexity which also constitute content.

ISO/IEC 12785-1:2009 defines *content* (item 3.7, in the meaning of *LET content* from a learning technology and content packaging point of view) as "*logical unit* to represent usable (and reusable) information contained in or related to learning, education, and training (LET) data in a formalized manner suitable for interpretation by human means". It is further explained by the "EXAMPLE: In the instructional context, content can be web-based instructional materials".

Thus, on the one hand a content entity is a logical unit representing *usable (and reusable) information* possibly contained in larger entities. On the other hand, "data in a formalized manner suitable for interpretation by human means" reveals that content "represents information" in a "manner suitable for interpretation by human means".

From the above it becomes clear that *data*, *information* and *content* convey different – though closely related – meanings although they are commonly used interchangeably. If "information is data that has been processed in such a way as to be meaningful to the person who receives it" (Riley 2012)[3], *content* becomes the 'representation of information meaningful to the person who receives it'. The latter comprises also the communication aspect – both in terms of technical as well as of human communication – which is particularly important in eLearning.

Today, there is a proliferation of content – not only in the Web at large, but also at organizations' level where it leads to high costs if it is not integrated in terms of *system integration* as well as *content integration*. When content is integrated in large organizations without being fully interoperable, it may become something called *big data*, which is not only just a great amount of content[4]. In this respect the Web can be considered as the biggest resource of *big data*.

If *content management* (according to ISO/IEC/IEEE 26511:2011, item 4.4) is "control of units of information with their metadata, to allow selective reuse in documents or information items with variable structures and formats", organizations are immediately faced with:

- The question of 'structure': structured content or unstructured content,
- The issue of complexity of content entities,
- The way of structuring information/content through data modeling using metadata.

3. Structured content

3.1. Structured content and unstructured content

Like data – information – content, *structured data – structured information – structured content* are commonly used interchangeably. This is due to the fact that "structure" is interpreted from different perspectives and in a variety of contexts. Thus content is approached and labeled as structured or unstructured according to the way it is interpreted or contextualized. This looks comparatively trivial when dealing with it in two well differentiated fields. However, difficulties emerge when ICT experts, marketing experts/consultants, webpage designers, content developers, editors and translators, among others, borrow and mix up terms without paying attention to the boundaries.

"Structured data can be defined as the data that resides in fixed fields within a record or file. Relational databases and spreadsheets are examples of structured data." (PC.COM s.a.) Apparently, unstructured data is the opposite, "data that does not reside in fixed locations. ... A huge amount of company information is unstructured text." (PC.COM s.a.) At the level of industry the need to cope with large amounts of *unstructured content* is evident:

> As companies increasingly create and store large amounts of information in electronic form, access to and the understanding of that information plays an important role in everyday business operations. However, much of the information that is generated and stored by companies is in unstructured form that is not suitable for either conventional relational database operations or for on-line analytical processing ("OLAP"). The unstructured content (e.g., e-mails, word processing documents, images, faxes, text files, Web pages, etc.) do not have any meaningful measure by which they can be compared with each other or combined to automatically communicate trends and/or abstract and diverse concepts ("attributes") that may be present across a number of types and/or categories of content.

> While some previous systems have attempted to classify and/or categorize unstructured content, such systems are generally rigid in nature and are not effective at measuring abstract and diverse concepts that span classifications and/or categories. Accordingly, there is a need for a measurement system or method for gleaning abstract and diverse concepts from unstructured content. (US Patent 7,249,312 B2 2007)

In the above, *unstructured content* largely refers to texts and other visual information/representation. This – in certain content management systems (CMS) – is extended towards auditory information (such as music) as well as other modalities and the respective media. The content entities here may be small or large. The metadata applied mostly refer to formal aspects, i.e. to syntactic structuring rather than *semantic structuring*. (See also: OASIS[5] s.a.)

Increasingly content demands a more refined *content management* that goes beyond the above-mentioned definitions from a technical point of view and includes semantic approaches. There have been lots of initiatives and developments dealing with structured and unstructured content in different sectors in order to overcome obstacles in the management of content. With the emergence of the Semantic Web and the need to adapt content from the Social Web demands with respect to semantic structuring are growing:

> The heterogeneous support content that is available for software products needs to be transformed to a semantically richer form in order to allow reasoning, adaptation and personalisation across it. … Semantic Web technologies such as ontologies represent an opportunity to base such structuring and markup on. The different types of content can be broadly categorised by their amount of existing metadata and structure. Consequently, different types of usage can be drawn from each: whereas highly structured content (such as technical documentation) can be used to derive an ontology of the knowledge domain, unstructured content (such as forum posts) can be marked up in order to provide querying users with a larger range of problem solutions. (Cena et al 2010)

In today's information society there is no content which is totally unstructured. However, *unstructured content* is not structured *sufficiently* from the point of view of semantic structuring.

In this connection *intelligent content* in the business sector (called or addressed as structured content) is in most cases unstructured content more or less semantically marked up. "Intelligent Content is structurally rich and semantically aware, and is therefore automatically discoverable, reusable, reconfigurable and adaptable." (The Rockley Group, Inc. 2008) In relation to *intelligent content*, Boses, in his article "Intelligent Content, Meet Content Intelligence", takes position against the unsuitable terms *structured data* and *unstructured data* and how the following two options can contribute to overcome the 'problem of naming' these concepts: "add structure and semantics to the data (Intelligent Content), or improve the technologies that try to understand 'unstructured' data". "When Intelligent Content and Content Intelligence work together the result is that unstructured data is transformed into meaningful and useful information that can support automation". (Boses 2012)

The difference between *structured content* and *unstructured content* seems to be determined by their amount and types of metadata and structure. *Entities of structured content* can be characterized by a high degree of semantic structuring based on the necessary metadata covering also the *semantic context* required to understand the content entity in question. It is typically processed and managed in databases which – if designed for textual data – should be capable of handling more than one language. *Entities of unstructured content* usually contain several/many entities of structured content in *co-text*. However, even if semantically tagged to a certain degree, unstructured content does not (maybe cannot or even should not) reveal the full semantic context of each entity of structured content contained. Besides, *unstructured content*, if textual, is usually monolingual (although it may contain elements in other languages).

There are many efforts and several approaches trying to overcome the gap between structured

and unstructured content by combining their features by means of complex CMS with several modules of different functions in order to make different kinds of content interoperable.

3.2. High or low complexity of content objects/entities/units

The complexity of a content entity in general could be determined in terms of:

- Quantity of information covered,
- Granularity of metadata applied,
- Amount of explicit context and co-text provided,
- Number of content types comprised,
- Degree of cognitive processing required, etc.

When developing short activities such as the ones found in podcasts, short videos, blog entries, tweets, short texts, wikis, eGames etc., complexity is usually viewed in terms of the information used. Here, the level of complexity in a way is governed by the length of the content, or reflects how much time and brain power is required for a person to understand and where appropriate do something, rather than simply memorise. (LIMBIC Learning Ltd. 2009, 2010)

Paradoxically, from the point of view of information processing:

> Highly structured knowledge bases permit a low degree of complexity to be managed by the information system. In contrast the degree of complexity is very high in weakly structured knowledge bases, whereby the user does only need a small amount of information about the meta-structure. (Zumpe & Esswein 2002)

Therefore, most approaches aiming at overcoming complexity due to weakly structured content are geared towards a higher degree of structuring while trying to avoid bothering the user with structure aspects at the user interface. This applies in fact to both, structured content and unstructured content.

Entities of structured content, as explained above, which are typically processed and managed in databases, are often called *microcontent* today. *Microcontent* "is a more general term indicating content that conveys one primary idea or concept, is accessible through a single definitive URL or permalink, and is appropriately written and formatted for presentation in email clients, web browsers, or on handheld devices as needed." (Dash 2002) *Microcontent entities* may comprise a day's weather forecast, the arrival and departure times for an airplane flight, an abstract from a long publication, or a single instant message. "Originally Jakob Nielsen (1998) referred to microcontent as small groups of words that can be skimmed by a person to get a clear idea of the content of a Web page. He included article headlines, page titles, subject lines and e-mail headings. Such phrases also may be taken out of context and displayed on a directory, search result page, bookmark list, etc."[6] The second use of the term (also called *microformats*[7]) extends toward other small information chunks that can stand alone or be used in a variety of contexts, including instant messages, blog posts, RSS feeds, and abstracts.

Comparing the approaches of terminological data modeling and the metadata-based approach of *microcontent/microformats* one can find many similarities. The whole Wikipedia is based on the *microformat approach* – however, not yet developed to a full *semantic* data modeling efficiency. When teaching a foreign language, lots of cultural, economic, historical, geographical and other facts – not to mention proper names – are important for learning a foreign language. This applies to common purpose language (CPL) as well as to special purpose languages (SPL – the object of teaching and research in LSP, language for specific purposes). It also applies to *linguistic entities* of structured content as well as to *non-linguistic entities*. Only fairly recently experts of terminology methodology started recognizing the importance of non-linguistic representations as well as of proper names in the field of specialized communication (in all its modalities – beyond spoken and written texts).

Microcontent – if taking into account terminological and lexicographical data based on the metadata approach – seems to be the most appropriate term representing the concept of *entities of structured content at the level of lexical semantics*. Therefore, in this contribution 'microcontent' will be used from here onwards.

The complexity of microcontent does not depend on the amount of information, nor on the technical formats, but can be determined in terms of:

- Granularity of semantics-oriented metadata,
- Degree of multilinguality and multimodality,
- Number of purposes for which they can be applied,
- Existence of cross-references to other entities within the same database or across repositories,
- Degree of cognitive processing required, etc.

From the above it becomes clear that the *complexity* of microcontent is governed by additional and partly substantially different criteria compared to those of unstructured content.

3.3. The way of structuring information/content through data modeling using metadata

Data modeling defines not just *data elements*[8], but also their structures and the relationships between them. Wikipedia[9] summarizes the information about data modeling as follows:

> *Data modeling* in software engineering is the process of creating a data model for an information system by applying formal data modeling methodologies and techniques. Its purpose is to manage data in a standard, consistent, predictable manner as a resource.
>
> The use of international data modeling standards is strongly recommended for all projects requiring a standard means of defining and analyzing data within an organization.
>
> Progressing from requirements to the actual database to be used for the information system the data requirements are initially recorded as a *conceptual data model* which is essentially a set of technology independent specifications about the data. The conceptual model is then translated into a *logical data model*, which documents structures of the data that can be implemented in databases. Implementation of one conceptual data model may require multiple logical data models. The last step in data modeling is transforming the logical data model to a *physical data model* that organizes the data into tables, and accounts for access, performance and storage details.

Several methodologies and techniques have been developed for the design of data models in order to guide data modelers in their work. However, two different people using the same methodology will often come up with very different results. Therefore, efforts are made to design generic data models which – being generalizations of conventional data models – "define standardized general relation types, together with the kinds of things that may be related by such a relation type." But as "the logical data structure of a database management system (DBMS), whether hierarchical, network, or relational, cannot totally satisfy the requirements for a conceptual definition of data ..., the need to define data from a conceptual view has led to the development of *semantic data modeling* techniques".[10]

State-of-the-art data modeling uses *metadata*, a term referring to 'data about data' which, however, is ambiguous, as it is used for two fundamentally different concepts:

- Structural metadata is about the design and specification of data structures and is more properly called 'data about the containers of data';

- Descriptive metadata, on the other hand, is about individual instances of application data.

ISO/TC 37 chose the second approach and is using standardized *data categories*[11] based on the international standard ISO 12620:2009 in order to assure the data exchange on the basis of the user´s need while paving the way for creating, extracting, combining, or adding data. The data categories (also called ISOcats) are registered and maintained in the Data Category Registry (DCR).

> ISO 12620:2009 provides a framework for defining data categories compliant with the ISO/IEC 11179 family of standards. According to this model, each data category is assigned a unique administrative identifier, together with information on the status or decision-making process associated with the data category. In addition, data category specifications in the DCR contain linguistic descriptions, such as data category definitions, statements of associated value domains, and examples. Data category specifications can be associated with a variety of data element names and with language-specific versions of definitions, names, value domains and other attributes. (ISOcat s.a.)

Considering content from the perspective of standardization in ISO/TC 37 as a continuum between terminological and lexical resources (considered as structured content here) and corpus resources (considered as unstructured content) demands a clear understanding and recognition of the entities that are labeled under the category of 'structured content'. *Data categories* are used to semantically structure *microcontent entities* (first of all terminological and lexicographical data) while the DCR permits to make clear the semantics of whole *microcontent resources*. This approach suitably fits the *microcontent* concept where the data and information must be enriched with metadata. However, these metadata should better be established based on the data categories' approach of ISO/TC 37.

The clarification above aims at contributing to:

- Promote re-usability of content – especially with respect to re-use in eLearning;
- Enhance the role of academia in content creation;
- Understand the value of distinguishing different types of content;
- Find the basis to harmonize methodological approaches for learning objects (LO);
- Bridge the gap between structured and unstructured content, such as:
  - Identify elements of structured content in unstructured content,
  - Insert/combine unstructured content in/with structured content,
  - Include increasing users' needs (different user/learner strategies, impairments, etc.);
- Use existing and emerging ICT technology in a more efficient way and based on sound methodologies;
- Identify gaps in standardization.

By means of terminological data modeling, the functional requirements for multilinguality, multimodality, multimedia, multi-channel output can be fulfilled from the outset. It also permits the improved use of unstructured content for the extraction of items of structured content as well as the use items of structured content for 'controlling' the quality, re-usability and interoperability of unstructured content. Furthermore, the didactic component can be added in case of designing LOs at the level of lexical semantics.

4. Microcontent

Microcontent (in the sense of *structured content at the level of lexical semantics*) indicates content that conveys one primary idea or concept. The kinds – not types – of objects/entities/items/units of this structured content may cover:

Designative concept representation	Descriptive concept representation	Possible extensions
(1) Terminological data:		
Linguistic designations:		
○ *terms* (incl. single-word and multi-word term) and similar, such a synonym, antonym, equivalent (in another language), etc. (written or spoken or other)	○ logic / partitive / other kind of determination[12] ○ logic / partitive / other kind of explanation ○ other kind of linguistic descriptive representation	○ *terminological phrasemes* (focused on LSP communication entities)
○ *abbreviated forms* (incl. initialisms, acronyms, clippings etc.) (written or spoken or other)		○ *terminological phrasemes* (comprising an abbreviated form)
○ *alphanumeric symbols* (written or spoken or other)		○ *terminological phrasemes* (comprising an alphanumeric symbol)
○ *proper names* (as kind of linguistic designation)		○ combinations of proper names with other linguistic designations
Non-linguistic designations:		
○ *graphical symbols*	○ graphical {descriptive[13]} representation (more or less systemic) ○ other kind of non-verbal descriptive representation (more or less systemic)	○ combinations of linguistic and non-linguistic/non-verbal designations
○ *other visual symbols* (incl. bar code, etc.)		
○ *non-visual non-linguistic symbols*		
		○ combinations of all kinds of designations
(2) Lexicographical data:		
○ *word* (or similar entities)	○ (different kinds of) explanations	
○ *collocations* (or similar entities)		○ *micro-utterances* (or similar entities)
○ *non-verbal communication entities*	○ (different kinds of) non-verbal explanatory representations	○ *entities of alternative and augmentative communication (AAC)*
○ *other kinds of entities of inter-human communication*		
		○ combinations of terminological and lexicographical data ○ combinations with entities of inter-human communication
(3) Controlled vocabularies:		
○ *thesaurus entries*	○ indications of conceptual structure as well as of domain/subject ○ other kinds of necessary indications	○ extensive mapping of controlled vocabularies
○ *classification entries*		
○ *entries of other kinds of controlled vocabularies*		
(4) Data categories (metadata):		
○ *names of data categories*	○ formal/coded description of data categories ○ additional (incl. non-formal/coded) elements of description	○ networking of repositories / registries of metadata and data categories
(5) Information of standardized and non-standardized coding systems for proper names and other entities in eApplications	○ coded entities (such as language codes, product classification codes, currency codes, etc.)	○ rules for the combination of these in various applications

Under a generic approach based on terminological data modeling methods all the above may be varieties of one data model for microcontent. If the development shows that entities like a day's weather forecast, the arrival and departure times for an airplane flight, an abstract from a long publication, instant messages, article headlines, page titles, subject lines, e-mail headings, blog posts, RSS feeds, etc. are becoming distinct types of microcontent, a typology can be established. If not the different kinds of microcontent could be dealt with in the form of a taxonomy, for instance.

Not to forget: complexity in the form of a higher granularity in terms of more (incl. more different kinds of) data categories in fact reduces complexity from the point of view of software engineering; a higher granularity of data categories does not necessarily indicate a higher degree of complexity.

5. Conclusions and outlook

Microcontent (in the sense of structured content at the level of lexical semantics) and unstructured content definitely are two types of content. Content management further distinguishes text, sound, image and video as content types and applies this also to *micrcontent entities* referring to units of reusable metadata that permits to create new 'information products'. Furthermore, a *content type* comprises a collection of elements; for classification of content, content types are chunks of meaningful information which are potentially content types but depend on business goals and needs of the users. Authors do not create content types; they create content items from content types; for example: strawberry is a content item that can be labeled under content type fruit, which in turn possesses content elements such as sweet and color. (Gibbon s.a.)

From the semantic point of view of terminologists this distinction of content types applied to microcontent is questioned. Inter-human communication occurs in the form of spoken, written and other kinds of communication. Under this perspective the different kinds of microcontent (and their different levels as well as combinations) can be classified in the form of a facetted classification scheme or taxonomy. From the point of view of a generic approach to the data modeling of structured content with the aim to achieve

- A generic data model (as described above),
- A semantic data model (possibly taking into account, but not necessarily following in all detail the existing approaches),

microcontent entities could be modeled based on one open-ended data model with a core structure (based on core data categories) and many variants depending on:

- The degree of multilinguality and multimodality/multimedia,
- The eApplication where used (at different level according to the target groups etc.).

Even information on sources (e.g. bibliographic and related data) can be part of this data model.

However, it is not impossible that certain kinds of *micocontent* on the Web – as mentioned above - may develop into content types in the future.

There two more push-factors for higher granularity in the direction of harmonized data modeling of microcontent:

(1) Legal issues in connection with copyright and other originators' rights are increasingly requiring even for minute pieces of information (even down to parts of fields)

- The need to reliably identify authorship (e.g. for the sake of allotting micro-credits);
- Securing the authenticity of entities of structured content (or parts thereof);
- Traceability of microcontent entities (or parts thereof) for the sake of assigning

rights and use conditions to re-use or re-purpose pieces of content in order to secure exploitation rights (whether asking for or waving remunerations), etc.

(2) The Recommendation on software and content development principles 2010 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 modeling (including the definition of metadata),

and hereafter throughout all the iterative development cycles. (MoU/MG 2012)

The above Recommendation inevitably requires a higher degree of structural complexity, which has to be coped with by a higher degree of data granularity of the data model. It may require additional URIs for the different language parts of the individual entries of structured content. Ultimately the time-honoured principle of *term autonomy* will have to be extended towards *representation autonomy* in the field of terminology management. This is anyhow necessary when re-purposing entities of structured content for instance for eLearning purposes. Because of the paramount phenomena of quasi-equivalence of concepts between languages links between parts of entries or even certain fields in a given entry to other entries (or parts thereof) will be necessary. As experienced in localization (such as in the field of technical documentation), the above even applies to non-linguistic representations due to cultural diversity factors.

6. Notes

[1] Data/information *objects, entities, items* or units are used interchangeably in experts discourse. Therefore, in this contribution 'entity' is used at the conceptual level, while 'entry' is used for the object/entity/item/unit in the data model.

[2] See for instance the *DIKW Pyramid* (also known variously as the 'DIKW Hierarchy', 'Wisdom Hierarchy', the 'Knowledge Hierarchy', the 'Information Hierarchy', or the 'Knowledge Pyramid'), which refers loosely to a class of models for representing purported structural and/or functional relationships between ***d***ata, ***i***nformation, ***k***nowledge, and ***w***isdom. "Typically information is defined in terms of data, knowledge in terms of information, and wisdom in terms of knowledge". (Rowley, 2007)

[3] "Note the two words highlighted in red – processed and meaningful. It is not enough for data simply to be processed. It has to be of use to someone – otherwise why bother?!" (Riley 2012)

[4] Webopedia (s.a.) refers to big data as "a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it is difficult to process using traditional database and software techniques"

[5] Advancing Open Standards for the Information Society

[6] See http://en.wikipedia.org/wiki/Microcontent#cite_note-1

[7] See http://en.wikipedia.org/wiki/Microformat

[8] Data element (in metadata standards according to ISO/IEC 11179-1:2004, item 3.3.8): unit of data for which the definition, identification, representation and value domain are specified by means of a set of attributes

[9] Wikipedia. Retrieved 2013-02-07: http://en.wikipedia.org/wiki/Data_modeling

[10] Wikipedia. Retrieved 2013-02-07: http://en.wikipedia.org/wiki/Data_modeling

[11] Data category (acc. to ISO 12620:2009, item 3.1.3): result of the specification of a given data field

[12] *Determination* according to Webster: in logic, the act of defining a notion [=concept] by adding differentia [=characteristics], and thus rendering it more definite. This corresponds also to similar use in physics <determination of nitrogen in the atmosphere> and in natural history <determination [=classification] determining the species of minerals, plants etc. to which they belong>)

[13] *Descriptive* with respect to non-verbal representations means that the representation indicates characteristics of the concept in question

7. References

ISO 5492:2008 Sensory analysis – Vocabulary

ISO 9241-151:2008 Ergonomics of human-system interaction – Part 151: Guidance on World Wide Web user interfaces

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/IEC 12785-1:2009 Information technology – Learning, education, and training – Content packaging – Part 1: Information model

ISO/IEC 15938-5:2003 Information technology Multimedia content description interface – Part5: Multimedia description schemes

ISO 24531:2013 Intelligent transport systems – System architecture, taxonomy and terminology – Using XML in ITS standards, data registries and data dictionaries

ISO/IEC 24800-3:2010 Information technology – JPSearch – Part 3: Query format ISO/IEC/IEEE 26511:2011 Systems and software engineering – Requirements for managers of user documentation

Boiko, B. (2004). Defining Data, Information, and Content. In B. Boiko, *Content Management Bible*, 2nd Edition, 3-12. Indianapolis: Wiley Publishing Company.

Boses, M. (2012). Intelligent Content, Meet Content Intelligence. Retrieved 2013-05-31, from contelligence.org: http://contelligence.org/intelligent-content-meet-content-intelligence-3/.

Cena, Federica; Dattolo, Antonina; Kleanthous, Styliani; Tasso, Carlo; Bueno V, David; Vassileva, Julita (Eds.). (2010). *Proceedings of the International Workshop on Adaptation in Social and Semantic. CEUR Workshop Proceedings,* Hawaii, USA. Retrieved 2013-07-05, from http://ceur-ws.org/Vol-590; http://sole.dimi.uniud.it/~antonina.dattolo/papers/2010/book/Dattolo-sasweb2010.pdf.

The Computer Language Company Inc. (1981-2013). Retrieved 2013-06-13 from Computer Desktop Encyclopedia: http://lookup.computerlanguage.com/host_app/search?cid=C999999&term=eContent&lookup.x=-410&lookup.y=-742.

Dash, A. (2002). Introducing the microcontent client. Retrieved 2013-07-05, from Anil Dash – A blog about making culture: http://dashes.com/anil/2002/11/introducing-microcontent-client.htmldictionary.

TheFreeDictionary (s.a.). Data. Retrieved 2013-06-30 from http://www.thefreedictionary.com/data.

Galinski, C. & Giraldo-Pérez, B. S. (2012). 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. In: *Terminologija* 19, 6-32

Gibbon, C. (s.a.). Content Modelling. Retrieved 2013-07-02, from Cleve Gibbon Homepage: http://www.clevegibbon.com/content-modeling/content-types/.
Gollner, J. (2013). A practical introduction to intelligent content. Retrieved 2013-06-30, from TC World: http://www.tcworld.info/rss/article/a-practical-introduction-to-intelligent-content/.

Gottlieb, S. (2008). Content is not Data. Retrieved 2013-06-30, from Content Here: http://www.contenthere.net/2008/05/content-is-not-data.html.

ISOcat (s.a.). ISO TC 37 Terminology and other language and content resources – Data Category Registry. Retrieved 2013-06-01, from ISOcat Homepage: http://www.isocat.org/.

LIMBIC Learning Ltd. (2009, 2010). Complex content. Retrieved 2013-07-26, from E-learning glossary a to z: http://www.limbiclearning.co.uk/resources/e-learning-glossary/a-to-z-listing/#complex-content.

MoU/MG-Management Group of the ITU-ISO-IEC-UN/ECE Memorandum of Understanding concerning eBusiness standardization) (2012). Recommendation on software and content development principles 2010. (MoU/MG/12 N 476 Rev.1) http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/MoU-MG/Moumg500.html.

Mullan, E. (2011). What is digital content. Retrieved 2013-05-13, from EContent: http://www.econtentmag.com/Articles/Resources/Defining -EContent/What-is-Digital-Content-79501.htm.

Nielsen, J. (1998). Articles/Microcontent: How to Write Headlines, Page Titles, and Subject Lines. Retrieved 2013-07-20, from Nielsen Norman Group: http://www.nngroup.com/articles/microcontent-how-to-write-headlines-page-titles-and-subject-lines/.

OASIS (s.a.). Retrieved 2013-07-01, from OASIS Unstructured Information Management Architecture (UIMA) TC: https://www.oasis-open.org/committees/uima/charter.php.

Öller, R. (2009). virtuelleschule.at. Retrieved 2013-06-13, from ViS:TV: http://www.virtuelleschule.at/vis-tv.

PC.COM (s.a.). Encyclopedia. Retrieved 2013-07-01, from PC.COM Homepage: http://www.pcmag.com/encyclopedia/term/52162/structured-data.

Riley, J. (2012). ICT, Business & Technology – Data and Information. Retrieved 2013-06-15, from Tutor2u: http://www.tutor2u.net/business/ict/intro_business_information.htm.

The Rockley Group, Inc. (2008).What is intelligent content. Retrieved 2013-05-30, from: http://www.rockley.com/articles/What%20is%20Intelligent%20Content.pdf.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. In: *Journal of Information Science 33 (2)*, 163–180

US Patent 7,249,312 B2 (2007) Attribute scoring for unstructured content. Inventors: Robert Joseph Jasper, Michael M. Meyer, Kelly Pennock

Webopedia (s.a.) Big data. Retrieved 2013-07-01, from Webopedia: http://www.webopedia.com/TERM/B/big_data.html.

Zumpe, S. & Esswein, W. (2002). Simplification of Knowledge Discovery using “Structure Classification”. In: W. Gaul, & G. E. Ritter, Classification, Automation, and New Media: *Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Passau, March 15–17, 2000*, 245-252. Dresden, Germany: Springer.