

Multimodal Common Format (MMCF) is an XML-based tag set to encode semantic structure of multimodal data. It is an extension of Global Document Annotation (GDA) tag set, which is to encode semantic structure of natural language documents. MMCF thus supports linkage between linguistic expressions or utterances and audio/visual data. Features of MMCF are discussed particularly in this respect, and an interactive multimodal presentation system using MMCF is described.

§1 Introduction

Language Integration Laboratory concerns integration of multimodal information by way of natural language, in order to build a software environment for not just development and evaluation of AI technologies but also their practical applications. We have accordingly developed **Multimodal Common Format** (MMCF), which is an insane of XML (eXtensible Markup Language)¹⁵⁾ to encode semantic structure of multimodal data possibly consisting of text, sound, and image. Many applications such as multimodal presentation and multimodal dialogue are expected to be composed by plugging various component technologies through MMCF, as depicted in **Fig.1**.

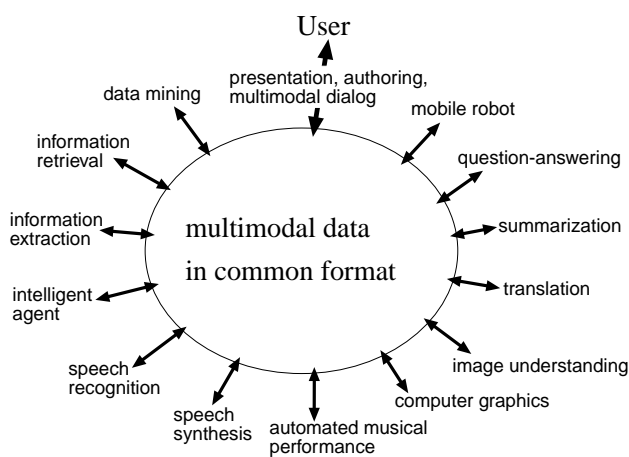


Fig.1 Information Integration with MMCF.

MMCF is an extension of the **Global Document Annotation** (GDA)⁵⁾ tag set⁶⁾. GDA is an initiative to spread a practice of annotating electronic documents to allow computers to recognize their semantic and pragmatic structures, drastically improving the accuracy and extending the functionality of machine translation, information retrieval, information filtering, data mining, consultation, expert system, and so forth. The GDA tag set has been designed to support such annotations. MMCF augments GDA with multimodal features. The notion of information integration in Fig. 1 is an extension of integration of NLP components^{4,9,1,16)}, which GDA aims at. GDA and MMCF aim at directly linking basic research and practical applications of intelligent systems, by having people semantically annotate a huge amount of data for practical purposes, as such data can also serve for basic research.

§2 Global Document Annotation

The GDA tag set has been designed by incorporating features of existing standards^{12,2,3,8)}. A new feature of the GDA tag set is that GDA-annotated texts are mapped to semantic networks. For instance, the data in **Fig.2** is mapped to the semantic network in **Fig.3**. XML elements (such as `<np>...</np>`) in sentences usually encode syntactic constituents. A `syn="bc"` means that

```

<su>
<persnamep opr="agt">Tom</persnamep>
<v sem="past.eng:meet">met</v>
<np opr="pat">
  <adp sem="sg">a</adp>
  <n id="g2" sem="eng:girl">girl</n>
</np>
<vp>
  <adp sem="ben">for</ad> <np eq="g2">whom</np></adp>
  <persnamep opt="agt">John</persnamep>
  <v sem="past.ont1:buy#2">bought</v>
  <np opr="pat">
    <adp sem="sg">a</adp>
    <n sem="deu:Blume">flower</n>
  </np>
</vp>
</np>
</su>
    
```

Fig.2 A GDA-annotated text corresponding to the semantic network in Fig.3.

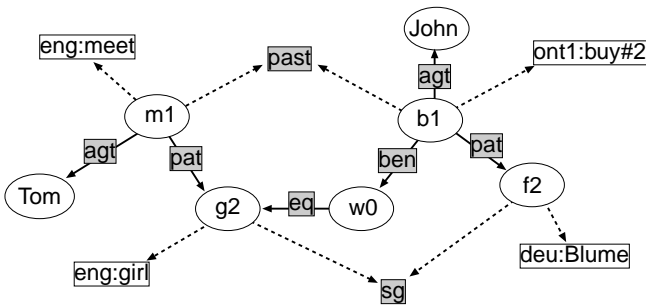


Fig.3 The semantic network encoding the meaning of "Tom met a girl for whom Bill bought a flower".

the child elements and texts form a chain of backward dependencies except that phrasal elements (<np> elements in the above example) cannot govern (be depended on by) other elements but just depend on the nearest non-phrasal elements. Thus in Fig.2 *Time* and *like an arrow* both depend on *flies*, and *an arrow* depends on *like*. *sem* attribute encodes word sense. In addition to the functionalities manifested in the current example, the GDA tag set also supports word senses, coreferences, scopes of various operators, rhetorical structures, and so on.

Fig.4 shows an authoring tool of GDA called tagging editor^{*1}, which allows users to rather directly manipulate XML annotation. Applications of GDA so far developed



Fig. 4 GDA Tagging Editor.

include summarization¹¹⁾ and slide presentation¹⁴⁾. GDA-based translation systems are under development, one of which uses UNL (universal networking language)¹³⁾ as the interlingua. Several linguistic corpora annotated with GDA tags are being developed as well⁷⁾.

§3 Multimodal Common Format

Text represents both the linguistic data (expression or utterance) and something else in the world that this data refers to. GDA treats text as such, having XML elements encode both linguistic data and their referents. MMCF extends this policy to all sorts of data, which include not only text but also image, sound, and so on. In MMCF, some XML elements (data elements discussed below) parallelly encode both these data and the things in the world they represent. This allows us to often conflate the data and their meaning in terms of this sort of elements, instead of redundantly encoding the structure of a data and the structure of its meaning.

The basic ontology of MMCF assumes the following three sorts of objects.

Semantic objects are what there are in the world, which are concepts, objects, events, and states of affairs.

Data objects are semantic objects of a special sort

*1 Available from <http://www.etl.go.jp/etl/nl/gda/TE/> together with another version of tagging editor and relevant documents.

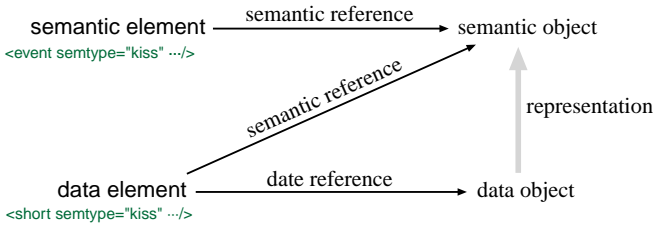


Fig. 5 Binary relations among elements.

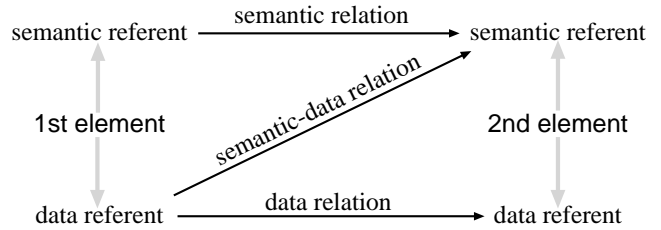


Fig. 7 Basic Ontology and Encoding in MMCF.

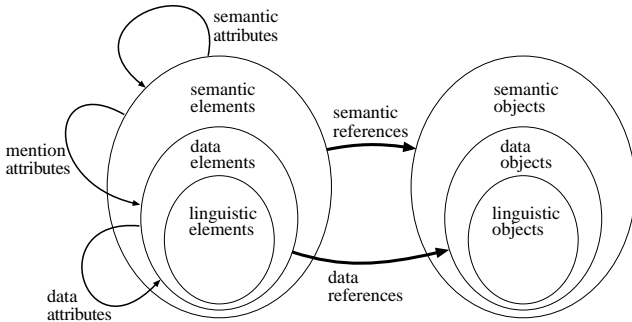


Fig. 6 Elements and their referents.

which are perceptual data such as image, sound, and text.

Linguistic objects are data objects of a special sort which are linguistic expressions (types) or utterances (tokens).

The three sorts of elements listed below are used to encode these objects.

Semantic elements encode semantic objects in general.

Data elements are semantic elements which encode data elements.

Linguistic elements are data elements which encode linguistic objects. Linguistic elements are defined in the GDA tag set, and MMCF extends the notion to that of data elements.

To define the notion of the **data referent** and

semantic referent of an MMCF element, let us postulate the following rules:

- Only data elements have data referent, and the data referent of a data element is the data object that it encodes.
- The semantic referent of a non-data semantic element is the semantic object that it encodes.
- The semantic referent of a data element is the semantic object represented by its data referent.

Namely, a data element encodes both its data referent and semantic referent, and a non-data semantic element encodes its semantic referent only, as summarized in Fig.5 and 6. For instance, linguistic (GDA) element `<np>a book</np>` encodes both noun phrase *a book* and the book it represents*2.

As shown in Fig.6 and 7, elements in MMCF can be linked with each other via the following IDREFS attributes*3:

Semantic attributes link between semantic elements, to represent semantic relationships (such as agent, patient, cause, concession, and so on) between their semantic referents.

Data attributes link between data elements, to represent syntactic relationships (such as part-whole relation) between their data referents.

Mention attributes link semantic elements to data

*2 Ambiguity arises here about the data referent and semantic referent of a data element. Namely, the data referent of 'a book' may be either the expression (type) or the utterance (token) of the noun phrase. Also, its semantic referent may be either the content of a book (type) or a particular print (token) of it.

*3 IDREFS attributes of XML are pointers to elements. We say such an attribute **a link** an element *X* to another element *Y* when *X* has an attribute instance `a="y"` where *y* is the id value of *Y*: that is, *Y* has an attribute instance `id="y"`.



Fig. 8 Correspondence between Utterance and Image.



Fig. 9 Annotation for Image Data.

elements, to represent semantic relationships of the semantic referents of the former with the data referents of the latter.

The name of a mention attributes is of the form $a.mt$, where a is a semantic attribute and $.mt$ is a postfix standing for *mention*, which specifies the data referent of the data element pointed by the mention attribute. Consider for example an event of someone's seeing some view. Then the relationship between this event and the view may be represented by mention attribute $obj.mt="V0"$ in a semantic element encoding that event, where $V0$ is the ID value of the data element whose data referent is that view.

§4 Authoring for Interactive Presentation

A version of MMCF is employed in an interactive multimodal presentation system developed in a joint project with Toyo Information Systems, Yokohama National University, and University of Tokyo, sponsored by Information-Technology Promotion Agency of Japan (IPA). The system has functionalities for authoring as well. Fig.8 shows how to annotate the correspondence between a text passage and a region of a 2D image. Fig.9

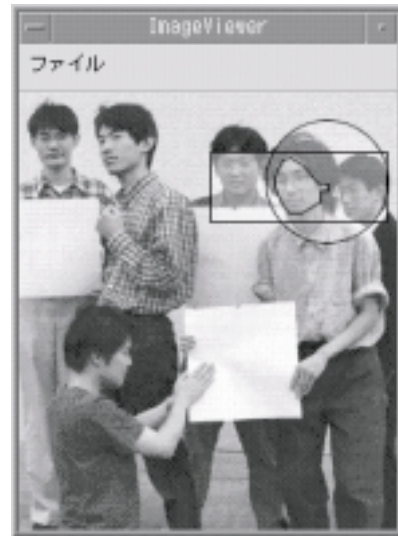


Fig. 10 Annotated Image.

displays the structure of a file annotating the image data in Fig.10. Note that the polygon, the rectangle, and the circle mentioned in Fig.9 are highlighted in Fig.10. Such regions can be manipulated from both interfaces. The current functionalities for interactive presentation are based on the technologies developed for text-based presentation¹⁴⁾. Further details of the system will be reported elsewhere.

§5 Concluding Remarks

MPEG-7¹⁰⁾ is another format to annotate multimodal data for semantic retrieval, among others, of multimedia contents. Its development is underway as a major activity of ISO. MPEG-7 as of June 2000 fails to address systematic treatment of data and semantic elements, but the above aspects of MMCF have been proposed to be implemented in MPEG-7.

References

- 1) H. Cunningham, K. Humphreys, R. Gaizauskas, and Y. Wilks. Software infrastructure for natural language processing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 1997.
- 2) EAGLES. *Expert Advisory Group on Language Engineering Standards*. <http://www.ilc.pi.cnr.it/EAGLES/home.html>.
- 3) Roger Garside, Steve Fligelstone, and Simon Botley. Discourse annotation: Anaphoric relations in corpora. In Roger Garside, Geoffrey Leech, and Anthony McEnery, editors, *Corpus Annotation — Linguistic Information from Computer Text Corpora*, pp. 66-84. Longman, 1997.
- 4) Ralph Grishman, editor. *Tipster Phase II Architecture Design Document*. New York University, NY, 1995. <http://www.tipster.org/arch.htm>.
- 5) Kôiti Hasida. Global document annotation. In *Natural Language Processing Pacific Rim Symposium '97*, 1997. <http://www.etl.go.jp/etl/nl/GDA/>.
- 6) Kôiti Hasida. *The GDA tag set*, 1998. <http://www.etl.go.jp/etl/nl/GDA/tagset.html>.
- 7) Kôiti Hasida, Hitoshi Isahara, Takenobu Tokunaga, Minako Hashimoto, Shiho Ogino, Wakako Kashino, Jun Toyoura, and Hironobu Takahashi. The RWC text databases. In *Proceedings of The First International Conference on Language Resource and Evaluation*, pp. 457-461, 1998.
- 8) Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313-330, 1993. <http://www ldc.upenn.edu/>.
- 9) D. McKelvie, C. Brew, and H. Thompson. Using SGML as a basis for data-intensive NLP. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 1997.
- 10) MPEG. MPEG-7 context and objectives, 1999. <http://drogo.cselt.stet.it/mpeg/standards/mpeg-7/mpeg-7.htm>.
- 11) Katashi Nagao and Kôiti Hasida. Automatic text summarization based on the global document annotation. In *Proceedings of the 17th International Conference on Computational Linguistics*, pp. 917-921, 1998.
- 12) C. M. Sperberg-McQueen and L. Burnard. *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. ACH, ACL, ALLC, 1994.
- 13) UNL Center. *UNL: Universal Networking Language — An Electronic Language for Communication, Understanding, and Collaboration*. Tokyo, 1996.
- 14) Masao Utiyama and Kôiti Hasida. Automatic slide presentation from semantically annotated documents. In *ACL'99 Workshop on Coreference and Its Applications*, 1999.
- 15) W3C. Extensible markup language, 1999. <http://www.w3c.org/XML/>.
- 16) Remi Zajac. An open distributed architecture for reuse and integration of heterogeneous NLP components. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 1997.

(Accepted June 30, 2000)

