An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language.

Mathieu MANGEOT-LEREBOURS

NII, Japan

mangeot@nii.ac.jp

Outline

- Context: From my Ph.D.
  - Accumulation of Lexical Resources
  - Existing Tools: SUBLIM, RECUPDIC & XML
- DML: Dictionary Markup Language
  - For New Resources, Generic
- CDM: Common Dictionary Markup
  - For Existing Resources
- Applications of DML/CDM
  - Consultation of Heterogeneous Resources
  - Online Edition of New Resources
- Conclusion

Accumulation of Lexical Resources

- At GETA/CLIPS Laboratory
  - MT dictionaries
    - Ariane MT System
    - UNL project
  - Human Usage Dictionaries
    - Ongoing Construction projects (Fe* projects)
- At XRCE Laboratory
  - Human Usage Dictionaries
    - Existing Resources: OHD, NODE, OES, ELRA
  - Resources for NLP (Morphological Analyzers)

Existing Tools & Methodologies

- G. Sérasset Ph.D: a Universal System for the Management of Multilingual Lexical Databases
  - Only theoretical, not implemented
- H. Doan-Nguyen Ph.D: a Methodology for the Recuperation of Existing Resources
- XML & Affiliates
  - XSLT, XSL, Xpointer, Xpath, Xlink,
  - XML Namespaces, XML Schemata

Dictionary Markup Language (1)

- Defines a Complete Framework for the Management of Lexical Databases
- Everything is described with an XML schema
- Namespace with a unique URI associated:

  http://www-clips.imag.fr/geta/services/dml
- Propose Notations to Define a Large Number of Microstructures: basic types, feature structures, trees, graphs, automata, functions, sets, etc.

Dictionary Markup Language (2)

Hierarchy of XML Elements described in the DML Schema:

- Lexical Database Data

  History, Users & Groups, Prefs & Profiles, API

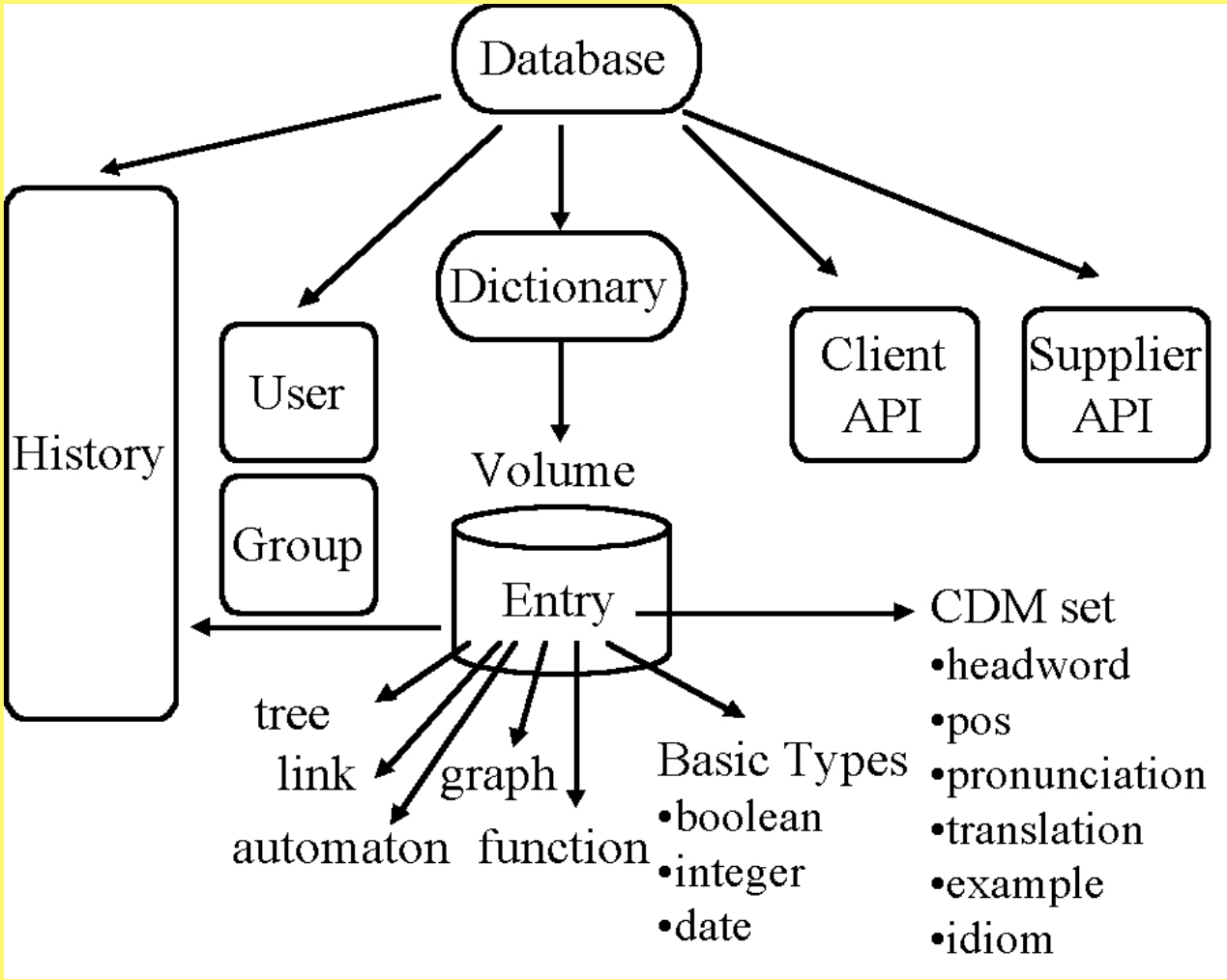
- Dictionary Metadata & Macrostructure

  Organisation & Links Between the Volumes

- Dictionary Microstructure (Generic)

  Structure of the Entries

General View of the DML



How To Manipulate Existing Heterogeneous Resources?

- Aim: Manipulating Heterogeneous Dictionaries without Modifying their Original Struncture and with Minimum Development
- Study of Existing Standards:
  - TEI, GENELEX, EAGLES, OLIF, etc.
  - Either too restrictive, or too complex

=> Creation of a Common Dictionary Markup

Common Dictionary Markup

- Set of Common Pointers Into Heterogeneous Existing Dictionary Structures
- Each Pointer Has a Unique Definition

<CDM elt>	**(tei equiv.)**	**<CDM elt>**	**(tei equiv.)**
<volume>		<translation>	(trans)(tr)
<entry>	(entry)	<example>	(eg)
<headword>	(hom)(orth)	<label>	(lbl)
<pos>	(pos)(subc)	<definition>	(def)
<pronunciation>	(pron)	<indicator>	(usg)

Applications: Edition & Consultation

- Online Edition with an XML Schema Compliant Editor
  - XML Spy, Morphon Java XML Editor, etc.
- Consultation of Heterogeneous Resources
  - DicoWeb: 10 Resources, 120 Users, 110 Req/Day
  - Papillon Project

    http://www.papillon-dictionary.org

Example of an Existing Volume

```
<dictionnaire name="DSZ"    date="26/04/99" source-language="fr" target-
<entree>
    <mot-vedette>affaire</mot-vedette>
    <administration><indexer date="Thu Nov 23 15:38:13 MET 2000">Mathieu
    <syntactic-cat><categorie>nom commun</categorie>
        <semantic-cat><traduction>ugy</traduction></semantic-cat>
    </syntactic-cat>
</entree>
<entree>
    <mot-vedette>aimer</mot-vedette>
    <administration><indexer date="Mon Jan 31 15:28:13 MET 2000">Mathieu
    <syntactic-cat><categorie>verbe</categorie>
        <semantic-cat><traduction>szeret</traduction></semantic-cat>
    </syntactic-cat>
</entree>
<entree>
    <mot-vedette>aller</mot-vedette>
    <administration><indexer date="Tue Apr 27 10:33:52 MET DST 1999">Matl
    <syntactic-cat><categorie>verbe</categorie>
        <semantic-cat><traduction>megy</traduction></semantic-cat>
    </syntactic-cat>
</entree>
```

Corresponding Metadata File

```
<volume-metadata xmlns="http://www-clips.imag.fr/geta/services/dml"
  xmlns:d="http://www-clips.imag.fr/geta/services/dml" xmlns:xlink="http:
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www-clips.imag.fr/geta/services/dml
  http://www-clips.imag.fr/geta/services/dml/dml.xsd" location="local" c
  00:00:00" hw-number="192460" installation-date="23/06/99 15:04:00" enc
  format="xml" name="Dsz_hun_fra" dbname="dszhunfra" version="1" source-l
  target-languages="fra">
  <authors>MML</authors>
- <cdm-elements>
    <cdm-volume element="dictionnaire" />
    <cdm-entry element="entree" />
    <cdm-entry-id element="entree" />
    <cdm-headword element="mot-vedette" />
    <cdm-pos element="categorie" />
    <cdm-translation d:lang="fra" element="traduction" />
    <cdm-corpus d:delay="10s" />
  </cdm-elements>
- <administrators>
    <user-ref name="Mathieu.Mangeot" />
  </administrators>
  <volume-ref xlink:href="dsz_hun_fra.xml" source-language="hun" />
</volume-metadata>
```

Conclusion

- Within the Papillon Project
  - Ongoing Work: Testing & Adjustement of the DML/CDM (Ask me for a Demo…)
- Within the Lexical Resources Community
  - Ongoing Work at ISO TC37/SC4
  - Needs for such an XML Markup Language