Dictionary for the XML Schema in CLANS

Shuai Wang
Department of Computer Science and Engineering
Chinese University of Hong Kong
wangs@cse.cuhk.edu.hk

1. Introduction

This is a dictionary for our proposed XML Data Schema in CLANS 2012. So far, it contains two entities, Person and Company. We have already well-defined plenty of nodes and attributes in use and reserved for future development. That is, current designed nodes and attributes are enough and proper for our current stored data, and we also define some potential and pre-prepared nodes and attributes for future-usage, like the <pages> and <media>, which are used for storing related links and media resources respectively.

Certainly, this schema would not be perfectly pre-designed for future development. However, since the XML Data Schema has its flexibility and extensibility properties, it is easy to improve and expanse our schema, like adding new nodes, attributes, or even new entities, with the project moving forward. Thus, we sincerely hope that this dictionary could to some extent help developers make use of current data and meanwhile, further extend this schema for the development of CLANS.

2. Defined Attributes

src: Source
This attribute is mostly attached to every element, since the value of this attribute demonstrates the origin of the text of the node. That is, it indicates where the information of this node derives from, so that if something wrong with the text content were detected, we can easily trace back to its original source. In this case we provide a robust approach to locate the cause of problem.

update: Update Time
The value of this attribute is a timestamp. This attribute is for recording the latest timestamp for updating, i.e., the time we make modification for a certain element.

desc: Description
The value of this attribute indicates the important and distinct feature the element belongs to.

pid: Personal Identification

This attribute is attached in the tag of <person> to identify which person the XML belongs to. The pid matches the personal key/index we define.

stock_code:
This attribute is for identifying a certain company with its stock code, in the tag of <company>, which is also the official identification of the listed company in stock market.

begin_time: Begin time for a record
This attribute is necessary for some tags, since it demonstrates the time when something happens. For example, the record (childNode) in timeline showed as followed:

```
<record begin_time="1990-00-00" end_time="1995-00-00">
    <company>交通银行股份有限公司</company>
    <dept>人事教育处重庆分行</dept>
    <position>处长</position>
</record>
```

end_time: End time for a record
The similar usage likes the *begin_time*.

3. Tag of node in Person XML file

(Please note that the listed tags are some ones that need further interpretation. Other tags that not displayed are considered the direct information it could be read from the tag name.)

person: A personal profile

pid: Personal ID, i.e., Personal Identification (key/index).

names: The parent node of the different child nodes for different kinds of names.

star_sign: The star sign. There are 12 options, like Gemini, Virgo, Aries.

birthplace: The place of birth, the place where a person was born.

pages: The related pages for one person, like Weibo, blog and homepage. It could be useful for our future crawled and extracted data.

current_jobs: The current jobs a certain person owns. It is understandable that a person may hold more than one job at the same time so under this <current_jobs> node is the childnode <current_job> that carries all related information. That is,

one <current_job> is a record for a particular job information.

ancestral_home: This is the place one's ancestors live, which is substantially different from the <birthplace>.

media: It is used for recording the path or URL and other related information like image and title, for related media resources.

timeline: It stores our preprocessed/extracted timeline information. It consists of records, and every one of them records one's working detail.

4. Values of the attributes inside the node of Person Entity

(The listed values are some ones that need further interpretations. Others that not displayed are considered the direct information it could be read from the attribute name.)

The *desc* attribute in <name> :
Sim_Chinese – Name in Simplified Chinese format
Tra_Chinese – Name in Traditional Chinese format
English – English name for one person
Nickname – Nickname for one person

The desc attribute in <address>:
Home – The living address for one person
Work – The working addressing for one person

5. Tag of node in Company XML file

(Please note that the listed tags are some ones that need further interpretation. Other tags that not displayed are considered the direct information it could be read from the tag name.)

stock_code: The identification code in stock market for a listed company.

company_names: The parent node of the different child nodes for different kinds of company names.

cross_code: The cross-code of Share A and Share B in Chinese Stock Market. This is a jargon.

industry_sectors: The sectors/categories that a company belongs to.

reg_add_longitude: The longitude of the registration address of a company.

reg_add_latitude: The latitude of the registration address of a company.

reg_staff_num: The staff number of a company in Registration.

reg_directors_num: The director number of a company in Registration.

reg_supervisors_num: The supervisor number of a company in Registration

offices_info: A company might own a headquarter and various branches. This node contains all related information about that.

6. Values of the attributes inside the node of Company Entity

The *desc* attribute in <company_name> :
Sim_Chinese – Company name in Simplified Chinese format.
Tra_Chinese – Company name in Traditional Chinese format.
Abbreviation – Abbreviation of a company name.

The desc attribute in <industry>:
General – The industry sector a company belongs to, with a general classification.
Specific – The specific sector a company belongs to, with a specific classification.

The desc attribute in <industry_code>:
General – The industry code corresponding to general industrial classification.
Specific – The industry code corresponding to specific industrial classification.