The Collaboration What Connects CTS and Database Data.

－From "a Classic Person's Name and a Character Chronological Table Database" Project－

AIDA,MITSURU
National Institute of Japanese Literature(NIJL)

Outline

CTS（Computerized Typesetting System）was developed for composition printing.We receive profits at various points by use of this system.For example, the following occurs.

(1)We can use the stable character set and perform data input.

(2)We can use a composition function together. By that, marking work can be done efficiently.

(3)We can process data easily...etc.

Thus, the outstanding performance is demonstrated when treating a lot of data and complicated structure.

1.Collaboration with the printing contractor.

If the data which should be inputted is extensive, it is more efficient to entrust the work to a special contractor.

Such as a character, a picture, and a sound...etc,there are many kinds of the data in which we can input commission.There are two kinds of contractors who treat the input of character data specially.

The first is a contractor who only inputs.And the second is a printing contractor.The latter had contracted only printing originally.However, recently, they contract an input by using an original system.

I will compare both feature.An input special contractor uses general-purpose apparatus and software.Therefore, this contractor is suitable when inputting the data of comparatively simple specification.In the data input of a small lot, there is an advantage of going up in many cases more economically than a printer.On the other hand, the printer can use a CTS (Computerized Typesetting System) system.Therefore, by complicated specification, it is possible to perform data formation covering many stages, and, moreover, a lot of data can be employed and managed stably.

There are few examples which choose a printer for the purpose of an input.However, in my database project, I have adopted the input using a printer's CTS system.And in order to manage by forming data efficiently, we (I and printer) came various devices and development.

2.What is the CTS?

CTS (Computerized Typesetting System) is made for Japanese composition.In order to create the newspaper of a voluminousness number immediately, its technical development was promoted.By development of a CTS system, a series of systems which perform collection of a manuscript, distribution, composition, and a space output were built.By using a computer, publication media acquired the speed and extensive distribution technology of the same communication of information as broadcast media.Moreover, it also became the automation by batch processing, and cultivating the way to database publishing further.This is the technology which progressed correction of the information produced frequently from increase in efficiency and the necessity for management.

This system was developed for the first time in Japan in 1970.And it has continued up to now, incorporating a function heterogeneous [the DTP system which progressed in the European-languages environment], and original, in order to make the rule thing of strict and complicated Japanese composition fully reflect.

At the beginning, many of CTS was the closed exclusive data environment.However, considerable open-ization has progressed in the past several years.Now, most problems are lost to compatibility with the data of the general business software of a personal computer.And it is decided also upon the document type definition file (DTD) by which compatibility with XML1.0 advised as a W3C standard was considered like "JepaX"(1999. 9).

Thus, CTS is becoming the existence opened gradually.The function extended in order to process a rule peculiar to Japanese composition, and specification are still in the state of exaggerated spec. to the thing used as standards, such as XML.

However, a pace of the technology which helps to advance an input, i.e., production of electronic contents, efficiently is loose, and the present condition is that even the input tool based on XML1.0 is not fully ready now.

3.The advantage of collaboration with the printer by CTS system use

There is a large scale database projects named "Japanes Classical Person and chronological Database[JCPD]" which is advanced in National Institute of Japanese Literature.This puts various basic historical records in a database about the person of Japan before modernization.There are various things from small-scale contents like "Heian-Jinbutsusi(平安人物志:The list of the peaple who lives in Kyoto)" and "Koudaiki(皇 代 記 :Empelor's Chronology)" to large quantity things like "Kugyou-Bunin(公卿補任:Court noble appointment table)" or "Sonpi-Bunmyaku（尊卑分脈: The compilation of genealogy）",but their all data is inputted and structurized for the database.Moreover, these data is linked with the picture of the person concerned.

We have completed the electronization of the data which surpass 30 kinds.The characteristic and the form of the data are various.Like genealogy historical records, if a special database system is not used, there are some which cannot fully demonstrate the characteristic of the data.（AIDA,MITSURU"The genealogy of the man who appears in the Japanese history, the research for a data base to turn"[Grant-in-Aid for Scientific Reseach(B),TERM OF PROJECT:(1998 ～ 2000),PROJECT NUMBER:10551013]）

This database project is carrying out the object of such data.And the reporter is advancing the project by collaboration with a printer, pulling out the potential power of a CTS system.

What kind of advantage is to perform input commission of data to the printer who has a CTS system? To it, the following three points can be considered.

(1)We can use the stable character set and perform data input.

(2)We can use a composition function together. By that, marking work can be done efficiently.

(3)We can process data easily...etc.

I want to describe each point hereafter, referring to an example.

3.1.The stable character set

The major company printer has achieved computerization of a printing system before JIS establishment.Therefore, there are not few companies which are applying by the original character set still now.The information adjustment with a JIS kanji code is secured by the conversion table defined for every user.

The character set used by the printer is specified to user real intention to the last.Therefore, the character thesaurus which systematizes a variant character is not created in a printer.An external character is created without any restriction by request of a user.

By the way, a JIS Chinese character is not what was not necessarily stabilized as everyone knows.It is not rare to produce type change occasionally by standard revision, either.

For example, the standard of a JIS character code-set has continued up to now through the following changes.

The JIS character code-set born for the first time is JIS C 6226-1978.The Chinese character code-set used those days this standard enacted, and every day was what is called "Chinese characters designated for daily use（当用漢字）"."Chinese characters designated for daily use" was changed into the "commonly used Chinese character(常用漢字)" in 1981.Since this revision was reflected, JIS C 6226-1983 was enacted in 1983.In order that the direction of a JIS standard might plan the adjustment of type then, it changed also into designs other than a commonly used Chinese character.The

criticism to this standard change was large.Then, when an auxiliary character code-set (JIS X 0212-1990) was enacted, the type of letters of the changed origin was also revitalized.JIS X 0212-1990 was abolished by establishment of Unicode 2.0 (JIS X 221-1995).However, the standard is still used.Then, at JIS X 208-1997, the cognition of the present type of letters was tried by introducing the concept of "inclusion."However, the reply which changes type other than a commonly used Chinese character into the thing according to the so-called type of "Kouki-Jiten(康熙字典)" was taken out with Council on National Language in 2000.Consequently, fundamental Japanese Chinese character sets will be scattered in JIS X 208/213/214.The character system of JIS X 208 had the meaning as the minimum character set in Japan by which private use is carried out generally and daily.But it is changed no less than 3 times in at least 20 years, and maintenance of the system also becomes impossible further.So it is hard to say that the JIS standard is fit for faithful electronic preservation of type information.The direction which commissioned the printer informational preservation can hold the type of original data stably for a long time.This is an ironical thing.

In our project, use proofreading work is done for the data outputted to paper by the font set of a printing office.In our project, the character set approximated to the type of the original is printed, and proofreading work is done.This is for preventing a proofreading person's capability, and the variation of judgment.

3.1.1 CTS←→JIS 3Steps

A conversion with the data and the exchange mark character for information (JIS character code-set) which were driven in for the character system for CTS is performed in the following ways.

①(Contractor) Reporting Characters without grounds.

First, the report for specifying conversion to the exchange mark character for information contained in the data which performed the input request is submitted.This is the character set used within TS system, and serves as a list of characters which cannot be performed if changing into an UCS2.0 Japanese side does not look for judgment.

②(Client)Creating Lists for Direction

As for conversion to the exchange mark character for information, directions are performed by judgment of an order person (here reporter).Now, a created conversion directions list is about 1,000 characters.

③(Both sides) Managing Characters which is not convertible

It is not avoided on the character of data that an external character occurs.And those

characters are used for a person and a name of a person in many cases.Therefore, , it is hard to perform substituting other characters.So, the "external character" is created in this project.This is for completing text information electronically.And it is because various text processing, such as creation of a corpus, becomes possible to all data.

These characters have many things which are not recorded on large-scale character sets, such as a "Konjaku-Mojikyou(今昔文字鏡)" and "GT Font-Style(GT 書体)", either.

3.2.Using a composition functions

The function which can be used by CTS composition -- for example, a side line, a kana, rate notes, and a table ..etc, there are many things.

The meaning of the information outputted as the appearance of the charge of financial funds being managed as a database is large.Each control code and tag are replaced by planned information, such as a tag of a database, and a delimiter, after an input end.Adoption of such a technique has greatly contributed also to the reduction of incidence to a data proofreading person, is the field which secures informational correctness and is raising the effect.

3.3.Text Processing

It is raised to the data which one composition control of the point which was excellent in the CTS system required that program processing is possible.Data processing of character substitution, extraction, sorting, etc. stabilized also to the large-scale data of hundreds of M bytes only in the text becomes possible.

For example, since the structure was complicated when we built a genealogy database, the input and formation work of data needed to be divided into several times of stages, and needed to be performed.There is the following in performed processing, for example in that case.

(1)Divide a record by making a punctuation into a delimiter.

(2)Extract date information and add A.D.

(3)Divide a record and add a management number to each.

In addition to this, the work which adds the information on reading was done using the reference table.

In recent years, PC is highly efficient.However, in addition, as for these processings, high load is required.Furthermore, by JIS, since there are various difficulties in processing of the data containing many characters which cannot finish being settled, it is still more so.

4.Conclusion

It seems that the project which inputs a lot of data steers a completely huge naval fleet.There, volitional understanding is clarified and it is necessary to ensure operation.

In case data input is begun, it is important to analyze a data structure appropriately in advance first.Next, the device for inputting and forming data efficiently and certainly is indispensable.When mobilizing especially a lot of people, the device which equalizes and transmits an input rule more correctly will also be important.The device which can perform error checking easily by minimizing judgment branch of each work process as much as possible in that case is desirable.

At the place of making data, the demand of enabling it to want to be able to perform such a device easily is serious.However, in the actual spot, there are many scenes against sauce data just like program language with which a lot of addition information was added.

This is a serious obstacle in respect of data sheet creation or database data proofreading.

We are continuing making the data of the scale of 2 million to 3 million characters every year.

The data sheet of the data format at the time of completion passed in that case is far.A data tag is entered in the copy of the original.The flag which had a structural system beforehand is entered in the data sheet.Regular database data is built by doing batch-processing work after an input end.

Generally, half a fault of database creation work is spent on making a data sheet and the verification work of data.And work pursuers fight with various addition information (tag marking) given to a text.

The structure language SGML was conceived from separating this information and appearance information on a document.The spotlight was captured as a language printing and for composition in the 1980s of the time of establishment.

Then, if it will enter in the 90s, the structure language will have made rapid progress as a data exchange protocol.Various things as which affinity with a database or a browser was considered have derived WML which specialized in HTML, XML, a cellular phone, etc., VRML (Virtual Reality Modeling Language) which aims at multimedia like 3D graphics.

Progress of the technology which, on the other hand, helps the input means corresponding to these formats is loose.

A CTS system has long history.However, it is hard to say that it is not necessarily made bearing database creation in mind.This is because it is the system which specialized in it pursuing an "expressional" device.Moreover, the original character set is adopted also about the character code, and the input based on a JIS standard character is not necessarily performed.

Now, downsizing is common.Under such a situation, the CTS system is truly heavy, thick, long and large.Therefore, it may be what is adopted only as a special company and a contractor.For a general use, a DTP system will replace for the role.In the meaning, it can be said that the input and the data construction technique like this

project are heterogeneous.

However, each know-how reported here is drawn out of a dialog with a printer.It is accumulated gradually, pulling out the potential power of a CTS system.

If this report can give you a help, it is pleased.