AQS: An Analytical Query System for Multi-Location Rice Evaluation Data

프란코 나자레노*, 정 승 현**, 강 유 진***, 이 경 희***, 조 완 섭***
(Franco Nazareno, Seung-Hyun Jung, Yu-Jin Kang,
Kyung-Hee Lee and Wan-Sup Cho)

Abstract Rice varietal information exchange is vital for agricultural experiments and trials. With the growing size of rice data gathered around the world, and numerous research and development achievements, the effective collection and convenient ways of data dissemination is an important aspect to be dealt with. The collection of this data is continuously worked out through various international cooperation and network programs. The problem in acquiring this information anytime anywhere is the new challenge faced by rice breeders, scientist and crop information specialists, in order to perform rapid analysis and obtain significant results in rice research, thus alleviating rice production. To address these constraints, we propose an Online Analytical Query System, a web query application to provide breeders and rice scientist around the world a fast web search engine for rice varieties, giving the users the freedom to choose from which trial it has been used, trait observation parameters as well as geographical or weather conditions, and location specifications. The application uses data warehouse techniques and OLAP for summarization of agricultural trials conducted, and statistical analysis in deriving outstanding varieties used in these trials, consolidated in an Model-View-Controller Web framework.

Key Words : Data Warehouse, OLAP, MVC, ANOVA in Augmented RCB

1. Introduction

Much rice varietal information has been collected over the years, and such collection has helped many researchers and scientists around the world. The technology of information and database systems helped in handling and managing these data, further alleviating the numerous investigations in improving rice. International cooperation and network programs in rice evaluation work hand in hand with national programs in providing such systems to help rice breeders and researchers around the world. Also, they conduct a partnership of annual evaluation of rice composition in pooling different rice varieties gathered from various varietal donations of cooperating countries, thereby promoting a free and global exchange of elite rice varieties and information. Within these networks and institutes, it is a must to have a systematic management of information and process in helping the team in continuously performing the various operations needed by the collection and evaluation program. Analysis and annual reports disseminated to the cooperating scientists for feedback are also important for succeeding experiments. This cycle of process lessens the burden of breeders performing their own varietal evaluation. However, with the

* Chungbuk National University, Dept. of Bio-Information Technology, 1st Author
** Chungbuk National University, Dept. of Information Industrial Engineering
*** Chungbuk National University, Dept. of MIS / u-Biz BK21 Team

voluminous data gathered, coping with the fast evolution of data propagation technology and information inquiries is one of the challenges faced. One solution is to provide an online query system accessible anytime through the World Wide Web. Such system can expedite the response to the cooperators' inquiries.

In this paper, we propose an online analytical query system using data warehouse and *OLAP* technology, combined with the statistical analysis used in *INGER*'s annual reports bind together in a *Model-View-Controller Model 2* web application framework.

2. Related Work

Within the International Rice Research Institute (IRRI), together with International Maize and Wheat Improvement Center (CIMMYT), rice information is centrally managed by *Crop Research Informatics Laboratory* through the International Crop Information System(ICIS), with specific rice implementation(IRIS). ICIS provides effective and consolidated global information on crop improvement and management both for individual crops and for farming systems[11]. Divisional databases are linked to ICIS central database and each are built with the same schema to allow ease of updates and uploads to the central database.

Together with Generation Challenge Program (GCP), the development of IRIS portal made varietal information on genetic resources and rice cultivars available online. This portal uses a special internet communication protocol and the GCP Domain Model[2] that expresses attributes, behavior and relationships between its entities, which is technology-independent. It provides users with searches for rice varieties, traits, evaluation sets and more from IRIS central database.

3. Online Analytical Rice Query System

3.1 Building the Data Warehouse

Data warehouses have the distinguishing characteristics which are mainly intended for decision-support applications such as OLAP and data mining techniques. They are optimized for data retrieval and analysis, not routine transaction processing. Data warehouses are designed precisely to support efficient extraction, processing and presentation for analytic and decision-making purposes. In this application, several tables in conjunction with the target data to be stored to the warehouse are considered. The trial observational data and trial information data are the two important tables to be considered, since they contain the actual data we want to exploit. <Table 1> shows the table data source for the constructed data warehouse using ETL.

<Table 1> Data sources for the new DW

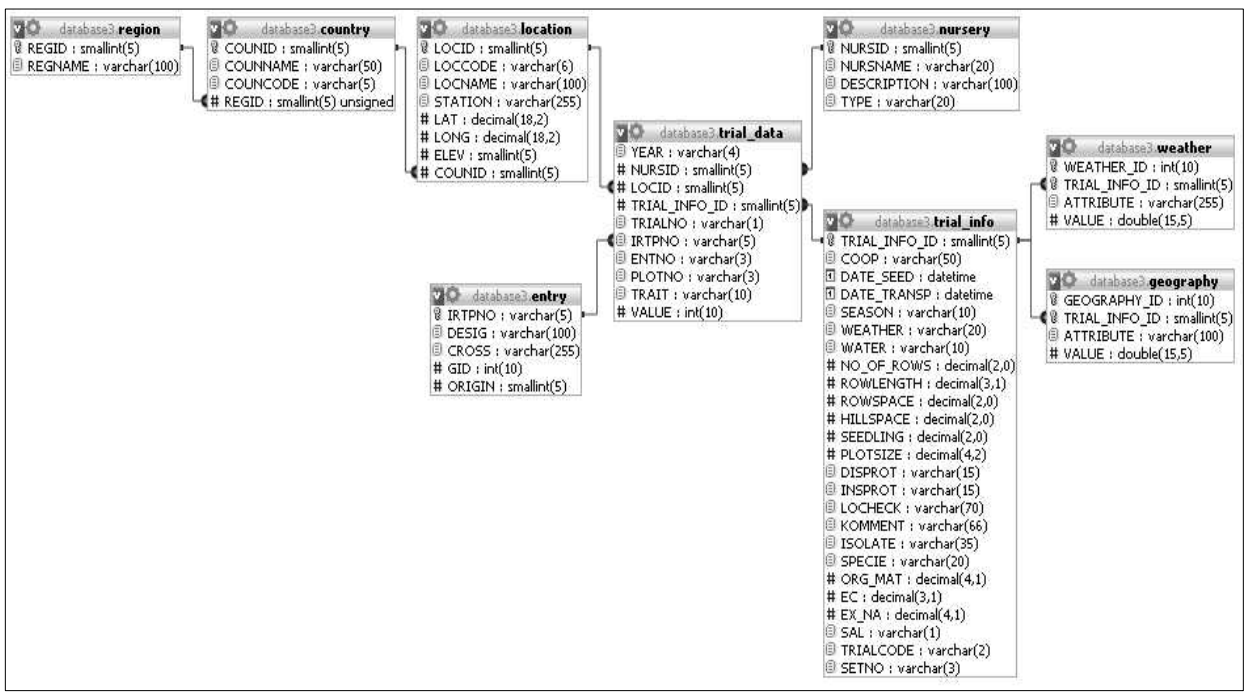
Table	Size(kb)	Description
Entry	2600	Contains the variety ID and name, along with information about its progenitors and origin.
Nursery	8	Contains the nursery ID and name with its description.
Country	13.5	Contains country ID and names who are participating in INGER trials.
Location	174.7	Contains the location ID, names and stations with geographic coordinate details.
Region	2.3	Contains the region ID and name according to INGER.
Nursent	15.6	Contains the final list of varieties used in the yearly nursery trials.
Geninfo	1300	Contains the information about the whole trial conducted in a specific location, nursery and year.
Nursery_data	147,700	Contains the actual recorded data of the observations during the trial period.

The extraction part is a basic ODBC connection to the Oracle database of INGER, allowing importing the tables in an MS Access database. This temporary database was used for the transformation and cleansing cycle. Steps in the transformation part includes translation of codes into its actual meaning (e.g. 1 = Highly Favorable, 2 = Favorable); data type conversion like measures in string data type are converted into integer or decimal type; Merging two columns like the country and location code are merged to form a representation of an unique location; Missing, null or zero values are handled carefully since different zero values can have different meanings depending on the context or trait observation. An example of this is that a 0 in grain yield observation means the variety did not yield during the trial, while a zero in a score for stem borer reaction means the trial experienced stem borer but did not inflict any damage to that variety. The cleansing part includes removal of undetermined data and suspicious values. As for the loading part, we acquire MySQL Migration Toolkit and we choose MySQL platform in storing the warehouse. It uses bulk transfers by turning off the constraints during the transfer and then turning on afterwards, checking for referential integrity constraints, e.g. foreign keys. <Figure 1> shows the resulting schema design of the new data warehouse.

With this resulting schema, several techniques can be used to deduce data, like rolling up and drilling down, as well as aggregations. A basic aggregate query that gives us the number of trials conducted by year, nursery and country is shown in <Listing 1>.

<Listing 1> Basic aggregate query

```
SELECT    a.year, b.nursname, d.counname,
          COUNT(a.trial_info_id)
FROM      trial_data a, nursery b,
          location c, country d
WHERE     a.nursid=b.nursid
AND       a.locid=c.locid
AND       c.counid=d.counid
GROUP BY  a.year, b.nursname, d.counname;
```



<Figure 1> New data warehouse schema

3.2 Model-View-Controller Framework

Following the MVC model 2 in implementing the online query system, we use *Spring Web framework* in *Java*, a popular open source application framework. The framework is a request-driven designed around a central servlet that dispatches requests to controllers with URL mapping. The views are implemented in *Java Server Pages* which are resolved by the Spring's View Resolver, View and SimpleUrlHandlerMapping interface classes. These are all declared as beans in the dispatcher servlet. As for the controller classes, it defines the basic responsibility of handling a request and returning a model and view. The model classes represent the domain, the services and the data access object for accessing the data warehouse objects. The service classes serve as the connecting object for the domain and DAO interface, and provide the services such as setters and getters. The DAO class is defined in a bean in the application context XML file. <Listing 2> shows the definition of the bean.

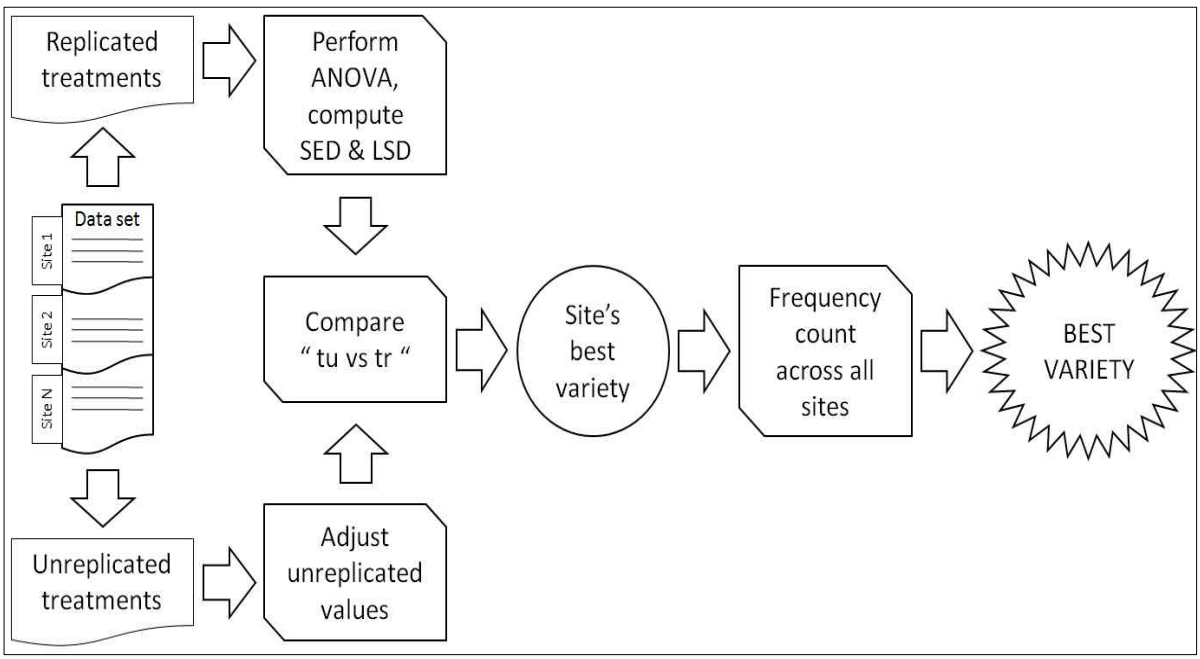
<Listing 2> Definition of DAO, Data Source and Property Configurer beans

```
<?xml version="1.0" encoding="UTF-8"?>
<beans xmlns= "...">
    <bean id="trialDao" class="..." >
        <property          name="dataSource"
                           ref="dataSource" />
    </bean>
    <bean id="dataSource" class="..." >
        <property name="driverClassName"

value="${jdbc.driverClassName}"/>
        <property name="url"
                  value="${jdbc.url}"/>
        <property name="username"
                  value="${jdbc.username}"/>
        <property name="password"

value="${jdbc.password}"/>
    </bean>
    <bean id="propertyConfigurer" class="...">
        <property name="locations">

<value>classpath...</value>
        </property>
    </bean>
</beans>
```



<Figure 2> Statistical analysis process in deriving the best varieties

3.3 Analysis of Variance for Augmented Designs

An augmented experiment design is useful for screening new treatments such as genotypes, insecticides, herbicides, drugs, etc. It is constructed by selecting an experiment design for the check treatments, and n number of new treatments, which can be in a Randomized Complete Block (RCB) or Incomplete Block (ICB) designs. INGER uses an Augmented RCB design for the trials, and the analysis are carried out using a two-way Analysis of Variance (ANOVA) comparing the cultivar means across locations[1], followed by multiple comparison procedure using least significant difference tests (LSD)[7]. First, the data set are retrieved from the data warehouse via the application, depending on the query parameters of the user. The resulting list is then passed on to the analysis engine to perform ANOVA and LSD comparison within the sites. The derived best varieties for each site are tallied and finally the engine outputs the resulting best varieties across test sites.

The datasets are divided into replicated varieties (*tr*) as checks and the unreplicated varieties(*tu*) as the test entries for each site. *Tr*'s are first computed with the usual ANOVA for RCB designs, ignoring the augments. <Table 2> shows the ANOVA table for the RCB used in analyzing the replicated treatments. The unreplicated treatments (*tu*) are adjusted for the block effects using the formula $Tadj = tu - AVG_i + GM$, where *tu* is the unadjusted value of the unreplicated treatments, AVG_i is the mean of the replicated treatment in the i-th replication, and *GM* is the overall mean of the replicated treatments. After adjusting the values, computations of the Standard Error of Differences (SEDs) and Least Significant Differences (LSDs) are done to be used for the comparisons. The SED to be used is for the difference between replicated treatment averages and unreplicated treatments, $\sqrt{(MS_E*(1+1/r+1/t-1/(r*t)))}$ where MS_E is the Mean Square Error, r is the number of replications and t is the number of treatments. The LSD is computed using the Student's t table forgetting the critical value from the t-distribution with Error(E) degrees of freedom and the computed SED. So, LSD is $SED*t(p,df_e)$, where p is the critical value and df_e is the Error degrees of freedom.

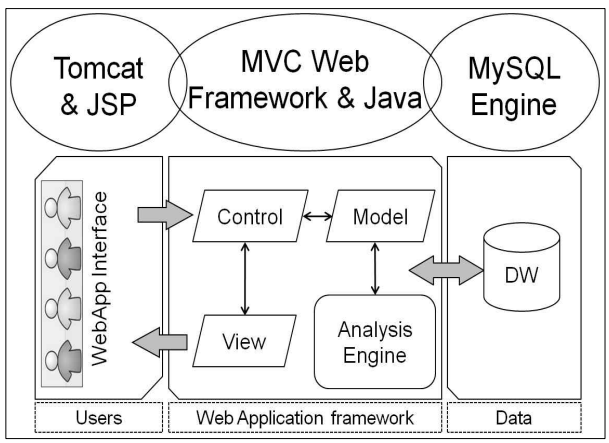
<Table 2> Analysis of variance as a RCB

Source	DF	SS	MS	Computed F-value
t	t-1	SS_t	SS_t/df_t	MS_t/SS_E
r	r-1	SS_r	SS_r/df_r	MS_r/SS_E
E	(t-1)*(r-1)	SS_E	SS_E/df_E	
Total	t*r-1	SS_{Total}		

The adjusted unreplicated treatments are tested against the better local check, thus if the difference between them are greater than the computed LSD, the treatment is significantly superior than the local check, otherwise it is considered significantly inferior.

3.4 Overall System Architecture

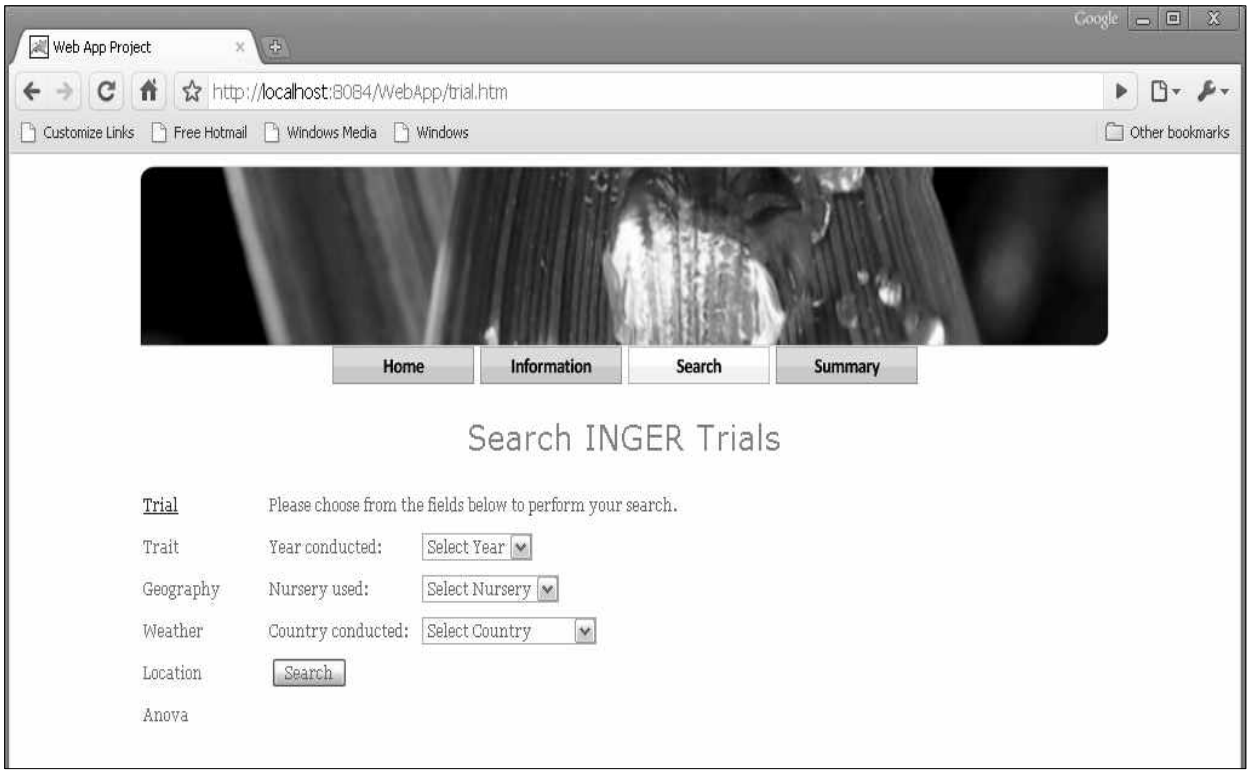
The online query system gives the user an instant result of the search and queries from the annual evaluation data. Using a web browser, the user chooses the parameters via the dropdown lists



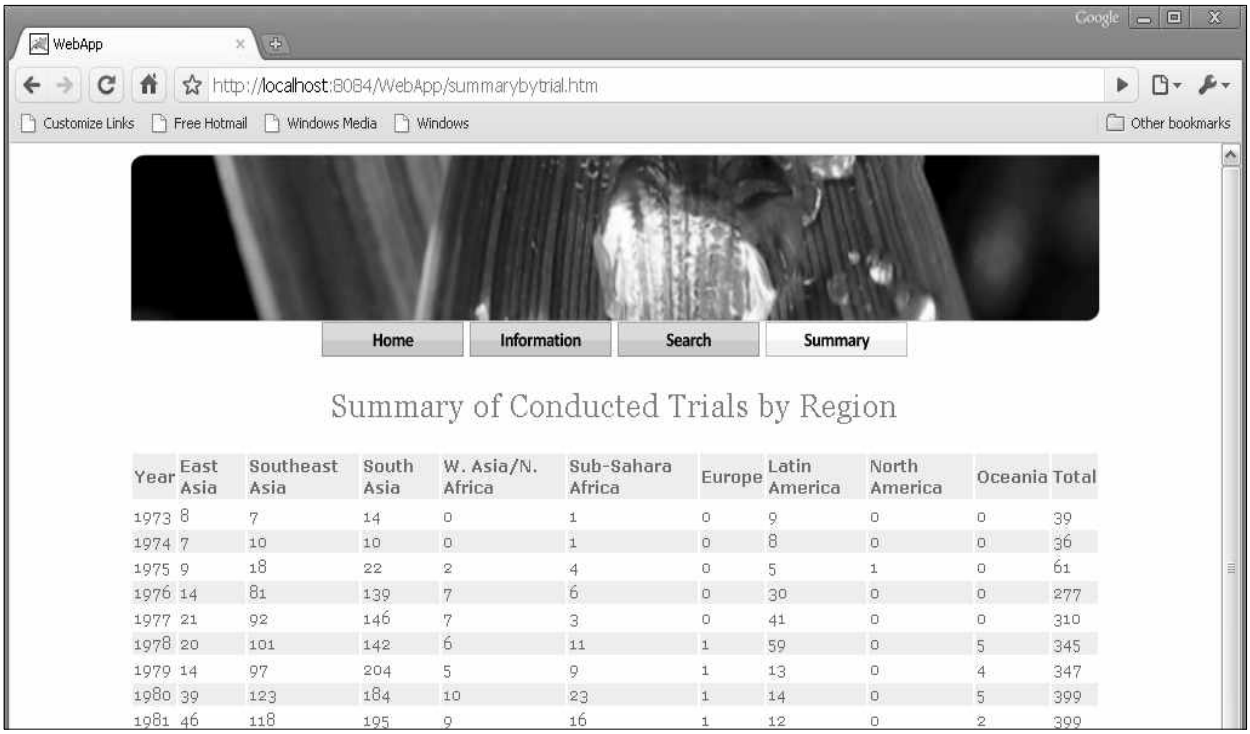
<Figure 3> Overall system architecture

which are populated from the data warehouse. A normal query will be a simple search for varieties used in the trials with factors of trait, geographical features, weather features and location or trial specification. The specific ANOVA search tool uses the analysis engine in deriving the best varieties for a given year, nursery and trait. <Figure 3> gives us the overall architecture of the online query system.

Users interact to the application through a web browser, in which parameters are provided to suit his search interest. The application's controller classes handle these inputs and commands and pass them to the appropriate model class, which in turn acquire data from the data warehouse. Once the model has the data, it will then perform the command specified, either OLAP or statistical analysis engine. The engine will then pass its result to the appropriate view page. A simple web design of the application for the query pages is shown in <Figure 4>. Users choose from the dropdown lists that suit his interests. The trial search shows the trial conducted for a year, nursery and country of choice; the trait search gives the varieties matching the chosen trait observation and specific value; for the geography and weather searches, users choose the recorded attributes for geography such as soil ph level and texture, fertilizer application, etc, while for the location search they can view the varieties used in a specific trial site, location and country. The summarization link let the users view the actual number of varieties tested in a year-nursery combination; summaries of number of conducted trials <Figure 5> and number of varieties used both grouped by region. The information link on the other hand gives the user the knowledge on the collected INGER varieties, nurseries and observational traits and attributes.



<Figure 4> Prototype web application



Year	East Asia	Southeast Asia	South Asia	W. Asia/N. Africa	Sub-Sahara Africa	Europe	Latin America	North America	Oceania	Total
1973	8	7	14	0	1	0	9	0	0	39
1974	7	10	10	0	1	0	8	0	0	36
1975	9	18	22	2	4	0	5	1	0	61
1976	14	81	139	7	6	0	30	0	0	277
1977	21	92	146	7	3	0	41	0	0	310
1978	20	101	142	6	11	1	59	0	5	345
1979	14	97	204	5	9	1	13	0	4	347
1980	39	123	184	10	23	1	14	0	5	399
1981	46	118	195	9	16	1	12	0	2	399

<Figure 5> Page for summary of conducted trials

4. Contribution

In this paper, we propose an online query system for agricultural evaluation data in promoting free exchange of varietal information. Specifically, a new data warehouse is constructed in handling and managing the data, and coping with the complexity of agricultural data. The data warehouse is constructed in such a way that it can accommodate rice information, general data about the conducted trials and more importantly the observed traits and reaction to stress or disease scores. This data warehouse is used by the underlying web application using the MVC web framework. The web application renders dynamic web pages giving the users a more flexible way to search for rice varieties suiting his needs and research interests. The web pages, using drop down lists, enables the users to look for varieties by selecting from the collected traits, overall geographical and weather conditions of the trials and location specification parameters. Within the resulting tabulated list, it links to more detailed information about the trials like scientists who conducted the trial, date of seeding and transplanting, season, etc., and more info about the varieties like pedigree information, country of origin, etc., giving the users a more concrete knowledge of his search. Also, outstanding varieties can be derived by using the ANOVA search function, indicating the year, nursery and trait parameter combinations. Finally, summarization of number of varieties used and number of trials conducted per year per nursery are provided to give the users the idea of how many varieties has been shared and used through this evaluation program. Also, users can find information about the varieties, the nurseries composed and trait description to gain knowledge of agricultural vocabulary and semantics.

Acknowledgement

This work was supported by the grant of the Korean Ministry of Education, Science and Technology (The Regional Core Research Program / Chungbuk BIT Research-Oriented University Consortium) and the Ministry of Education, Science and Technology (MEST) and Korea Industrial Technology Foundation (KOTEF) through the Human Resource Training Project for Regional Innovation. Special acknowledgements to IRRI-INGER for providing the source database and collaboration efforts in the statistical analysis implemented for this application, particularly to Dr. Edilberto Redoña, Ms. Ma. Concepcion Toledo, Ms. Nadine Singson and Mr. Victor Alcantara.

References

[1] J. P. Bradley, K. H. Knittle, A. F. Troyer, "Statistical methods in seed corn product selection," Journal of Production Agriculture, vol. 1, pp.34-38, 1988.

[2] R. Bruskiewich, et. al., "Generation Challenge Program (GCP): Standards for Crop Data," OMICS Journal of Integrative Biology, vol. 10, no. 2, pp.215-219, 2006.

[3] R. Elmasri and S. Navathe. Fundamentals of Database Systems. Boston, MA: Addison Wesley, 2006.

[4] W. Federer, M. Reynolds, J. Crossa, "Combining results from augmented designs over sites," Agronomy Journal, vol. 93, pp.389-395, 2001.

[5] K. Gomez and A. Gomez. Statistical procedures for agricultural research. New York, NY: Wiley-Interscience, 1984.

[6] J. J. Johnson, J. R. Alldredge, S. E. Ullrich, "Replacement of replications with additional locations for grain sorghum cultival evaluation," Crop Science, vol. 32, pp.43-46, 1992.

[7] R. Kimball and M. Ross. The Data Warehouse Toolkit: The complete guide to dimensional modeling. New York, NY: Wiley Computer Publishing, 2002.

[8] S. Ladd, K. Donald, D. Davison, S. Devijver, C. Yates. Expert Spring MVC and Web Flows. Berkeley, CA, 2006.

[9] J. Machacek, A. Vukotic, A. Chakraborty, J. Ditt. Pro Spring 2.5. Berkeley, CA: Apress, 2008.

[10] C. G. McLaren, R. Bruskiewich, A. Portugal, A. Cosico, "The International Rice Information System - A Platform for Meta-Analysis of Rice Crop Data," Plant Physiology, vol. 139, pp. 637-642, 2005.

[11] The Spring Framework - Reference Documentation. Retrieved on February, 2010 from http://www.springsource.org/.

프란코 나자레노 (Franco Nazareno)

• 2008.9~현재 충북대학교 바이오정보기술학과 바이오인포메틱스 전공 석사 재학중

• 관심분야: Database Systems and Data Warehouse, Object-Oriented Programming, Web Applications and Services, Bioinformatics Tools and Solutions

정 승 현 (Seung-Hyun Jung)

• 2007.2 충북대학교 정보산업공학과 석사
• 2007~현재 충북대학교 정보산업공학과 박사 재학중

• 관심분야 : 바이오인포메틱스, 데이터웨어하우스, 클라우드 컴퓨팅

강 유 진 (Yu-Jin Kang)

• 2007.2 충북대학교 경영정보학과 석사
• 2007~현재 충북대학교 경영정보학과 박사 재학중

• 관심분야 : OLAP, Business Intellegence, 데이터웨어하우스

이 경 희 (Kyung-Hee Lee)

• 2004.2 충북대학교 전자계산학과 박사
• 2009~현재 충북대학교 기업정보화센터 박사후과정생

• 관심분야 : OLAP, 데이터웨어하우스, Business Intellegence, 네트워크기술, ERP

조 완 섭 (Wan-Sup Cho)

• 2005.11~현재 충북대학교 경영정보학과 교수
• 관심분야: 바이오인포메틱스, 데이터웨어하우스, OLAP, 데이터베이스

논 문 접 수 일 : 2010년 06월 08일
1차수정완료일 : 2010년 06월 19일
게 재 확 정 일 : 2010년 06월 20일

본 논문은 2010년 한국산업정보학회 춘계학술대회에서 우수논문상을 수상하였으며, 한국산업정보학회논문지 편집위원회의 심사과정을 거쳐 본 호에 게재됨.