The Genomic Plant Warehouse Framework:

A Systematic Literature Review

Teddy Siswanto
Information System Department, Trisakti University
Doctor of Computer Science
BINUS Graduate Program, Bina Nusantara University
Jakarta, Indonesia
teddysiswanto@trisakti.ac.id

Nesti Fronika Sianipar
Food Technology, Faculty of Engineering
Research Interest Group Food Biotechnology
Bina Nusantara University, Jakarta, Indonesia

Bahtiar Saleh Abbas
Doctor of Computer Science
BINUS Graduate Program, Bina Nusantara University
Jakarta, Indonesia

Spits Warnars Harco Leslie Hendric, Harjanto Prabowo
Doctor of Computer Science
BINUS Graduate Program, Bina Nusantara University
Jakarta, Indonesia 11480

Meyliana
Information System Department
School of Information Systems, Bina Nusantara University
Jakarta, Indonesia 11480

Achmad Nizar Hidayanto
Faculty of Computer Science
University of Indonesia
Depok, Indonesia

***Abstract*—Designing data warehouse of genomic plant is different compared to business data warehouse. Due to the development of biological science today that leads to other scientific field such as bio-technology, bio-chemistry, bio-science and bio-informatics. For that required framework design in accordance with the development of scientific genomic plants. This paper uses a literature review approach to research published in 2002-2017. The reference chosen for the systematic literature review are EBSCO, Science Direct, Emerald Insight and Proquest. Through literature research has been tested as many as 77 papers from all sources related to the criteria, then obtained as study candidate of 30 papers based on title and abstract relevance based on research questions. After further investigation, there are only 19 papers that can be used in this paper. The result shows that the variable of framework for the genome data warehouse plant are Construction & Content and Search, that component consisting of Data Source, Data Model, Data Integration, Visualization, Data Schema, Data Mining, DNA Extraction, Web Services, Data Projection, Data Template-Import Tool-Web Interface and Display Format,-Filter-Further Analysis. While search component obtained search Sequence, Keyword and Query Builder, Region, Unstructured Keyword and Structured Text Search.**

Keywords—Framework; Data Warehouse; Construction; Content; Search

I. Introduction

Refers to article contained in the Republika online media on January 18, 2017 which states "Ribuan Tanaman Herbal di Indonesia Belum Optimal Dimanfaatkan"[21], it shows concern for Indonesia's unmanaged natural wealth. On that basis came the idea to build a data warehouse for the development of plants. Plants are a type of unstructured data that is different from the business data through the normalization process can form a type of structured data that is relative easily stored in the database repository.

In researching in the field of bioinformatics, the research of plant genomes is very important.[1] Data Warehouse will monitor the progress of sequencing and maintenance of information quality standards easily. In developing standard plant annotation approaches requires a series of automated analysis tools and platforms that allow regular annotations. The Data Warehouse system differs in design and capability, as well as the intended users. In this case as an example, a system can provide generic and configurable solutions easily, while other systems allow biologists to customize data warehouses to their specific needs. A system provides a simple text field for query in the database query, while other for advanced users, the system can provide a more convenient search tool designed. Scientists who do genomic curation work have been trained in scientific disciplines such as biology and computer science. Curation of genome is a process digitizing and integrating separate different genomic data and related literature to facilitate the sharing of genome knowledge [20]. The use of standardized terminologies and metadata schemas can facilitate the process of genome curation. The use of multiple data mining tools in biological research has made the data more intensive and evaluation process. Scientists use a data-driven approach to conduct general research practice[8]. To manipulate curation and data, equipment is required in accordance with its specific biological context by scientists[19]. Biologists in their research also require the support of data analysis from computer experts to process the massive data set generated, but the task is not easy to implement because the traditions

978-1-5386-2930-7/17/$31.00 ©2017 IEEE

and culture of the domain cannot be the same. Genome scientists need closer scrutiny to explain the characteristics of domain knowledge in both computer science and biology. Through these exams they can apply effective practices to ensure data exchange and data quality among various disciplines [11]. Therefore, this research will to define **"The Framework How to Design Genomic Data Warehouse?"** Because in addition to warehouse must be able to store unstructured and structured data. The purpose of the research is to find out the appropriate framework component of the genomic plant data warehouse so that the database design can store the plant DNA data and can access to obtain the required information and knowledge.

II. Methodology

This research conducts a thorough study of literature studies on research on the genomic plant warehouse framework. The process is grouped into sections, namely: determining the source of research, to determine keyword patterns for the search process in the database through initiation of inclusion criteria and exclusion criteria, data mining and analysis of findings to answer research questions. [25]

A. Searching Process

The initial step determines the source of the literature to find the appropriate article or journal. The sources selected for the systematic literature review are :

o	EBSCO	(www.search.ebscohost.com)
o	Science Direct	(www.sciencedirect.com)
o	Emerald Insight	(www.emeraldinsight.com)
o	Proquest	(www.proquest.com)

The applied keyword pattern for finding research papers related to answering research questions is made using the Boolean operator to filter the data, so that we can specify the priority for searching data based on the symbols used. Boolean symbols and operators that use in this research such as OR, AND. The combination of these keywords is:

— (genomic OR DNA) AND plant AND warehouse AND framework

The search mechanism inclusion criterion consists of three filter processes. The first process is "Studies Found". Any document found from the source publication associated with the specified keyword will be saved as a Published Foundation. The next step the paper be filter by title and abstract. If the title and abstract are match and complimentary to determine the research question, those papers will be stored as "Candidate Studies". The final step for filtering papers are all documents included in the candidate will be read thoroughly to answer research questions. If the paper matches the research question, those papers will be defined as "Selected Studies". While to clarify the validity of the literature, the search exclusion criteria are defined in several procedures, namely:

- Papers based on the date of their publication prior after 2000
- All identities journal or conference, author's identity are mentioned on complete paper structure.
- Duplicate paper from the same study will not be included in the Systematic Literature Review

B. Data Extractions

The research literature has examined 77 papers from all sources and criteria. Of the 77 papers studied, there are 30 papers that study candidates based on related titles and abstracts for research questions. After further investigation, there are only 19 papers that can be used in this study.

TABLE I. Data Extraction in Inclusion Criteria

Sources	Found	Candidate	Selected
EBSCO	17	6	6
Science Direct	29	9	5
Emerald Insight	4	2	1
Proquest	27	13	7
Total	77	30	19

III. RESULTS AND DISCUSSIONS

This study aims to determine the components of data warehouse framework. The purpose of the establishment of a data warehouse is to transmit data and simultaneously as knowledge discovery, which is to search the relationship between information with each other that has not been known at the time the information is stored. The data is stored in a very large number of databases. This required information technology to process the data rapidly. In this section, this paper presents the trending characteristics of the Selected Studies literature, such as publication sources, publication years, variable component classifications, and warehouse mapping frameworks from the study literature. On Table 2, it shows title, year, type and journal (J) or conference (C) name.

TABLE II. Source of Publications

No	Title	Year	Journal-J/ Conference-C
1	TargetMine, an… [4]	2011	J
2	TRUNCATULIX.. [10]	2009	J
3	INDIGO– INte … [1]	2013	J
4	OPTIMAS-DW: … [5]	2012	J
5	Systems Integration .. [18]	2009	J
6	BiNA: A Visual …. [7]	2014	J
7	Genomic Data Mo… [12]	2002	J
8	*In Silico* Analysis… [3]	2008	J
9	Biological work …. [6]	2004	J
10	A MapReduce ….. [13]	2016	J
11	An ontology-……. [16]	2011	J
12	Geminivirus data .. [22]	2017	J
13	Plant systems ….. [23]	2014	C
14	Domain knowledge.. [11]	2014	J

15	Database and Tools .. [17]	2014	J
16	Data integration … [14]	2015	J
17	The Metagenomics .. [24]	2016	J
18	Global catalogue …. [15]	2013	J
19	Interoperability of … [9]	2005	J

The most of authors discipline expertise come from Multi Discipline (53%), Computer Science (32%), Biology (10%) and one paper from Consortium (5%). For multi disciplines come from Bio Sciences, Bio Technology, Bio Chemistry and Bio Informatics, while for computer science comes from School of IS, School of IT and Computer Science, while for Biology comes from Micro Biology and Biology as seen in the following figure.

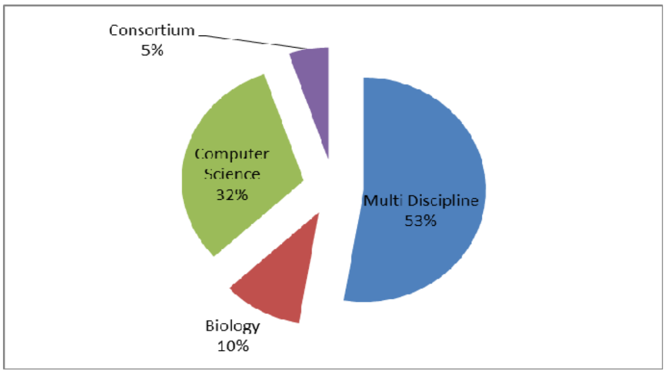

Fig. 1. Graph of Authors Discipline of Expertise

TABLE III. THE COMPONENT DATA WAREHOUSE

Variable	Component	Indicator
CONSTRUCTI-ON AND CONTENT	Data Source #(4)	The definition of the entity's genomic data model to be stored [4,15,22,23]
	Data Model #(2)	New data models such as integrated protein DNA interactions in new ways [4,12]
	Data Integration #(4)	Extract, transform and load (ETL) are used for integration and import of data[10,12,14,22]
	Visualization #(4)	Each component of the visualization model has a set of properties, such as shape, size, label text, text font, color, line width [7,22,23,24]
	Database Schema #(3)	The database schema consists of about 60 tables for storing raw data as well as meta information and analysis result [1,5,10]
	Data Mining #(2)	The genome sequences and complete ORFs are classified using Machine Learning approaches [22,23]
	DNA Extraction/ ETL #(1)	Deploying and testing Genomic (DNA) and bioinformatics standards will help improve methods in the Meta Genomics field [24]
	Web Services #(1)	Interoperability and compatibility of resources remains one of the biggest challenges for bioinformatics [9]
	Data Template, Import Tool Web Interface#(1)	OPTIMAS-DW supports an innovative concept for connecting data from different data domain thru meta data [5]
	Data Projection #(1)	The generic projection framework supports mapping of arbitrary external information [7]
	Display Format, Filter and Further Analysis	The catalog information of each collection, used to build the global catalog database framework. GCM workflow scheme. [15]
SEARCH	Sequence	The search and analysis tools provide various searching criteria on both nucleotide sequences [22]
	Keyword and Query Builder Search	The keyword search option provides a simple interface to the underlying data annotation. Query builder, giving more control over annotation classes and attributes sought, restricted, or viewed. It is possible to combine several questions through logic constraints [1]
	Region Search	Territorial search used to know the characteristics of a particular genome area. [1]
	Unstructured Keyword and Structured Text Search	Search words or sentences can be combined with a Boolean operator [18]. Keywords used in search are unstructured, structure and text keywords.

Based on the components identified and then described in a framework as follows :

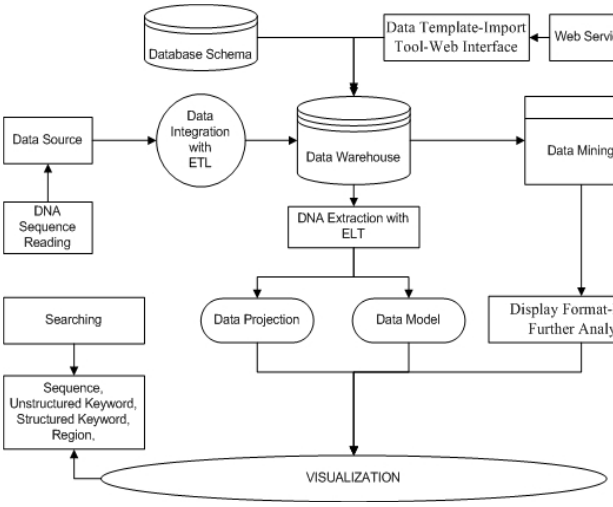
Database Schema
Data Template-Import Tool-Web Interface
Web Services
Data Source
Data Integration with ETL
Data Warehouse
Data Mining
DNA Sequence Reading
DNA Extraction with ELT
Searching
Data Projection
Data Model
Display Format-Filter-Further Analysis
Sequence, Unstructured Keyword, Structured Keyword, Region,
VISUALIZATION

Fig. 2. The Framework of Genomic Plant Data Warehouse

Data Warehouse Schema is formed based on Data Source obtained from DNA Sequence Reading entered through Extract, Transformation and Loading (ETL) process. Data Warehouse will also be able to receive data source from Open Access other Data Warehouse through Web Service and Data Templates-Tool Import-Web Interface. For the development of hidden data discovery and further analysis will be continued through Data Mining. In order to display Visualization according to information needs and knowledge changes, the data will be extracted by Extract, Loading and Transformation (ELT) process using Data Projection and Data Model. Search data, information and knowledge can be done by using the keywords Sequence, Unstructured, Structured and Region.

IV. IMPLICATION AND CONCLUSION

The result of Systematic Literature Review in 19 papers, obtained the similarity of component framework that researched more than one that is Data Source, Data Model, Data Integration, Visualization, Database Schema, Data Mining, while the specific result of research author in the form of DNA Extraction, Web Services, Data Projection, Data Template-Import Tool-Web Interface and Display Format-Filter-Further Analysis. These are summarized in the Construction & Content and Search variables.

Of the 19 papers studied, many of its authors come from multidisciplinary of 53%, indicating that scientific development now requires the cooperation of researchers in various disciplines such as bio science, bio technology, bio chemistry and bio informatics.

V. LIMITATION & FUTURE RESEARCH

Based on the component of framework is identified, there are many emerging areas to be considered for future research. The result components only a conceptual components framework for genomic plant warehouse and there are many aspects of component framework to be refined. It has been a challenge to organize the component framework, while there are many theories to support it but the number of database is restricted, so the amount of the papers limited to represent the fact completely. Therefore it needs extensive empirical testing to validate those components. The result of the framework component in data warehouse found may be increased if the data source is expanded and also the specified range of year can be extended. For future, research can be developed thru identify the components in DNA sequence reading to determine the data warehouse scheme is the most appropriate and test the process flow from and to the genomic plant data warehouse.

REFERENCES

[1] Alam I. Antunes A. Kamau AA. Ba alawi W. Kalkatawi M. et al, INDIGO – INtegrated Data Warehouse of MIcrobial GenOmes with Examples from the Red Sea Extremophiles, PLoS ONE 8(12): e82210, 2013

[2] Antonio B.A. et al, Rice at The Forefront of Plant Genome Informatics, Genome Informatics 11, 1-11, 2000

[3] Chen Ming and Harrison Andrew, *In Silico* Analysis of Crop Science: Report on the First China-UK Workshop on Chips, Computers and Crops, Genomics, Proteomics & Bioinformatics, Volume 6, Issues 3–4, Pages 190–19, 2008

[4] Chen Y-A. Tripathi LP. Mizuguchi K, TargetMine an Integrated Data Warehouse for Candidate Gene Prioritisation and Target Discovery, PLoSONE 6(3): e17844, 2011

[5] Colmsee C. Mascher M. Czauderna T. Hartmann A. Schluter U et al, OPTIMAS-DW: a comprehensive transcriptomics, metabolomics, ionomics, proteomics and phenomics data resource for maize, BMC Plant Biol 12: 245, 2012

[6] Farmerie, W., Hammer, J., Liu, L., Sahni, A. & Scneider, M. Biological workflow with BlastQuest, Data Knowledge Engineer 53. 75–97, 2005

[7] Gerasch A., Faber D., Kuntzer J., Niermann P., Kohlbacher O., Lenhof H.-P., Kaufmann M.: BiNA: a visual analytics tool for biological network data, PLoS ONE 9. 2. e87397, 2014

[8] Goth, G, "Preserving digital data", Communications of the ACM, Vol. 55 No. 4, 2012

[9] Hiten Vyas Ron Summers, "Interoperability of bioinformatics resources", VINE, Vol. 35 Iss 3 pp. 132-139, 2005

[10] Henckel K, Runte KJ, Bekel T, Dondrup M, Jakobi T, Ku¨ ster H, Goesmann A, TRUNCATULIX-a Data Warehouse for The Legume Community, BMC Plant Biol 9: 19, 2009

[11] Huang, Hong. "Domain Knowledge and Data Quality Perceptions in Genome Curation Work." Journal of Documentation 71.1: 116-42, 2015

[12] Jake Yue Chen and John V. Carlis. Genomic Data Modeling. Information Systems, 28(4):287-310, 2002.

[13] Kamal S, Ripon SH, Dey N, Ashour AS, Santhi V. A MapReduce approach to diminish imbalance parameters for big deoxyribonucleic acid dataset. Comput Methods Programs Biomed 131:191–206, 2016

[14] Lapatas V., Stefanidakis M., Jimenez R. C., Via A. and Schneider M. V. Data Integration in Biological Research: an overview, *J. Biol. Res.* **22**. 9, 2015

[15] LinhuanW., Qinglan S.,Hideaki S. ,Song Y.,Yuguang Z.,Kevin M.,AlexanderV. , Suzuki k.I.,Moriya O.,Yeonhee L.et al. Global catalogue of microorganisms (GCM): a comprehensive database and information retrieval, analysis, and Visualization System for Microbial Resources, *BMC Genomics* 14:933, 2013

[16] Li, Y.-F.; Kennedy, G.; Ngoran, F.; Wu, P.; and Hunter, J. An ontology-centric architecture for extensible scientific data management systems. Future Gener. Comput. Syst. 29(2):641–653, 2013

[17] L.S. Jing, F.F.M. Shah, M.S. Mohamad, N.L. Hamran, A.H.M. Salleh, S. Deris, et al., Database and tools for metabolic network analysis, Biotechnol. Bioprocess Eng. 19 568–585, 2014

[18] McGarvey, P. B., Huang, H., Mazumder, R., Zhang, J. et al., Systems integration of biodefense omics data for analysis of pathogen-host interactions and identification of potential targets., PLoS One 4 e7162, 2009

[19] Pruitt, K.D., Tatusova, T., Brown, G.R. and Maglott, D.R. "NCBI reference sequences (RefSeq): current status, new features and genome annotation policy", Nucleic Acids Research, Vol. 40 No. D1, 2012

[20] Reed, J.L., Famili, I., Thiele, I. and Palsson, B.O., “Towards multidimensional genome annotation”, Nature Reviews Genetics, Vol. 7 No. 2, 2006

[21] Republika Online. “Ribuan Tanaman Herbal di Indonesia Belum Optimal Dimanfaatkan", http://nasional.republika.co.id/berita/nasional/daerah/17/01/18/ojyked359-ribuan-tanaman-herbal-di-indonesia-belum-optimal-dimanfaatkan, Accessed July 17, 2017

[22] Silva et al. Geminivirus Data Warehouse: a Database Enriched with Machine Learning Approaches, BMC Bioinformatics 18:240, 2017

[23] Sheth BP, Thaker VS, Plant systems biology: insights, advances and challenges. Planta 240:33–54, 2014

[24] The MetaSUB International Consortium, The Metagenomics and Metadesign of the Subways and Urban Biomes, *Microbiome*, 24(4), pp.1–14, 2016

[25] Meyliana, Achmad Nizar Hidayanto, and Eko K. Budiardjo. "The critical success factors for customer relationship management implementation: a systematic literature review." *International Journal of Business Information Systems* 23.2 (2016): 131-174.