
Ice Hockey Research Data Platform from Official Records Data and Verification

Seung-kyo Jin*, Ji-hyun Jang**, Hye-young Kim***, Sun-tae Kim****

ARTICLE INFO

Article history:

Received 30 September 2019

Revised 28 November 2019

Accepted 12 December 2019

Keywords:

Ice Hockey,
Official Record Data,
Research Data,
PDF Parser,
Ice Hockey Reference

ABSTRACT

In this study, a database was established by analyzing the record data research produced in ice hockey sports. The deployed data verification with Ice hockey reference service was demonstrated with ice hockey officials and players. This research utilized the data stored in the KNSU Datanest data repository and developed PDF parsers for batch processing of records. Among the types of records, the game summary, team roster, team statistics, and player statistics files were collected, and tables were extracted from the records. PDF records were converted to text in CSV format which are converted to DataFrame and loaded into the database. Out of the total 22 types of records, 4 types were constructed with OO data parsed as element values. Data verification has found no problems with the quality of the data deployed, showing a high satisfaction with providing 66 factors against the 30 factors provided by the service previously used.

1. Introduction

With the fourth industrial revolution, the “Web of Documents” era is changing to the “Web of Data” era. The value of data is large enough to be compared to crude oil. Collecting necessary data efficiently and adding value to the data is to secure knowledge-base. Efforts to solve global challenges by utilizing secured knowledge-base are being actively researched in web, data science and artificial intelligence subject matter. These studies are changing our society and human lives. Researchers spend a lot of time reading literature and comparing all existing approaches (Yu et al., 2019). In particular, they spend a lot of time analyzing data to support data-driven decision making and to create evidence to support policy decisions. Through this process, the importance of data

* Master's student, University of Science & Technology (Korea Institute of Science and Technology Information, KISTI), Republic of Korea (t4716@kisti.re.kr) (First Author)

** Software planner, Argonet, Republic of Korea (balance@argonet.co.kr) (Co-Author)

*** Professor, Korea National Sport University (KNSU), Seoul, Republic of Korea (hykim@knsu.ac.kr) (Co-Author)

**** Assistant professor, Jeonbuk National University, Republic of Korea (kim.suntae@jbnu.ac.kr) (Corresponding Author)

International Journal of Knowledge Content Development & Technology, 9(4): 31-45, 2019.
<http://dx.doi.org/10.5865/IJKCT.2019.9.4.031>

is recognized. Data analysis and application of various sports sites such as baseball, soccer, volleyball, and basketball are also becoming common. Ice hockey researches are also active to meet these needs. However, the current application of artificial intelligence in the sports sector is only applied to reading raw data. Only reading raw data within a single device is applied. Until now, artificial intelligence in winter sports has not produced as much innovative results as in areas such as image processing, voice recognition, and natural language processing. However, the importance of field data is recognized in the sports and sensor studies are introduced to collect site data. In other words, the technology development in the collection phase is taking place and the analysis of the data and application of the site are not yet available as expected. At this stage, the most important thing to apply artificial intelligence to the sports is to collect and manage data through the platform. All data from equipment or facilities must be collected on the appropriate platform to facilitate connection with artificial intelligence. Therefore, it is necessary to establish an integrated sports data management system to boost data-based research in the ice hockey. This requires an understanding of sports data. Winter sports experimental data are produced in the research process. It should be provided to reuse data a variety of context information as well as simple description information. Therefore, the development of metadata elements and schema design are needed to accommodate data produced in winter sports experiments. In addition, to support efficient data analysis and reuse, raw data needs to be loaded into the system at the component value level.

2. Literature Review

Yu et al. (2019) carried out a study in which the researchers automatically extracted and analyzed the stored information in tabular form in a paper written in PDF so that they could use the research time efficiently. To automatically read PDF tables, rule-based and learning-based functions were used, and the tabula-java libraries used in this study were used in this process. In order to boost tourism, which is important for Sri Lanka's economy, Mihindukulasooriya, Priyatna, and Rico (2018) extracted tourism statistics data present in PDF files and converted it into 5-star Linked Open Data in a study that wanted to open tourism statistics data to the public for use in intelligent business decisions. To extract tables from the PDF files, the tabula-java library was used. Lee (2010) suggested the integrated sports archive considering each types of sports so that preserve and manage the records to improve its value. Research used by ice hockey official records was conducted. Koo et al. (2016) analyzed 2014-2015 season of Asia League official records to get important features in ice hockey using logistic regression. This research indicates that analyzing the official match records can help shape national team strategies and plan training. If there was longer period of dataset, research could be expanded the scope of study. Systematic management of official records was needed to improve performance of ice hockey player and promote the research through ice hockey record data. Jin, Kim, and Jang (2018) suggested classifying the record elements present in ice hockey games into match information, offense, defense, face-off, team information and playing time information, to design and store the schema in the database. We could create ice hockey research ecosystem by parsing and storing official ice hockey records and providing them in report

form for statistical analysis. We looked forward to establishing the process of archiving and using recorded data. In future research, we expected to expand the scope of data to unstructured data as represented in the review of big data technology in professional soccer by Memmert and Rein. They suggested possible machine learning model through tracking data, medical data and physiological data etc.

3. Methodology

In this study, PDF formatted raw data from the International Ice Hockey Federation (IIHF), a research data sharing platform developed by the Korea Institute of Science and Technology Information (KISTI), was used. A parser has been developed that can load raw data produced in the ice hockey field at the component value level. Schema was designed to efficiently utilize the loaded element values, and statistical websites were established to verify the quality of the data deployed. Data quality was verified by demonstration of statistical websites to athletes and leaders at ice hockey sites. In the preceding study, the schema designed by Jin, Kim, and Jang (2018) was extended and reflected in the database. Data collected from the IIHF and stored in Datanest were imported in batches. Parsers for bringing in various types of data in batches have been developed. Transformation to CSV format is required to extract tables from raw data in PDF format. At this time, the rule base was used to solve a problem where the row and column of the table did not match. Out of a total of 22 types of records, parsers were developed for four categories (Game summary, Team roster, Team statistics, Player statistics). The data quality has been verified by providing an interface to experts at the sports site that provides the deployed data in statistical form. Finally, the interview was conducted with Korean women's ice hockey officials to draw up information to be reflected in the follow-up study. Figure 1 below shows the Batch Processing System (BPS) and Web Service System (WSS) developed in this study. In BPS, PDF conversion, data extraction, and data entry are performed on PDFs managed by Datanest. WSS provides the values of elements loaded into the database through various user interfaces.

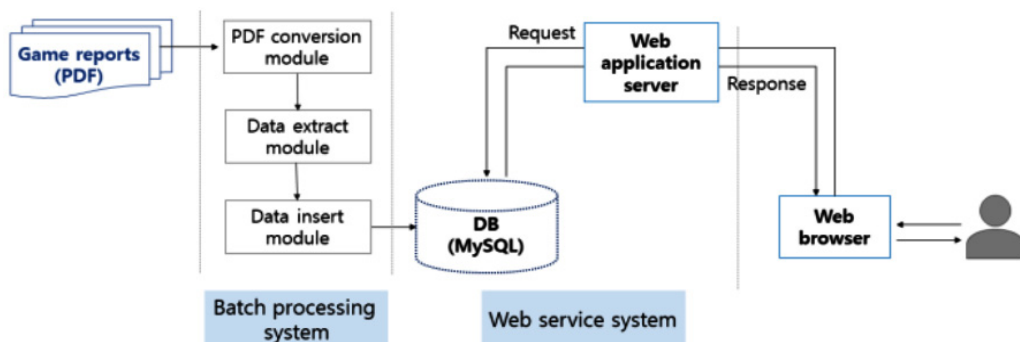


Fig. 1. Flowchart of whole system

4. Development and Quality Assessment

Data collection was carried out through IIHF. The archive of the IIHF official website (www.iihf.com) provides information on all competitions held on the IIHF. The sites were organized separately by season/league, so we downloaded statistical data in PDF format from the Statistics category within the corresponding league site. The folder structure was established according to the type of IIHF site, and official statistics were collected by season, gender and league and mounted in the Datanest repository in a compressed file format. The IIHF World Championship Record registered in the system is the most detailed description of the ice hockey competition record. Factors such as pace-off and power play scoring probability are included in the sports arena. The world championships men's first division game summary record includes actual time on ice for each competitor, which is important information for assessing their performance. Since competitions will be held for national athletes, figures that can be applied to the Olympics can be applied. For the World Championships, it is the largest number of international competitions. It is possible to apply for analytical research using ice hockey data from the most participating countries, but it is difficult to load additional data from various competitions due to different forms from Asian leagues, Asian Winter Games and domestic records. Metadata elements were derived, and databases were designed to reflect these characteristics.

4.1 Database Design and Data Deployment

The Game summary record is available in five table formats, which include meta information, team's records, player's records, and event occurrence for each competition. The data was stored in seven tables according to the data presentation form. A gameInfo table containing the hosting competition information was extracted from the match's meta-information. Team statistics table represents the data of the team's periodical records. Period tables are separated by goal, penalty and penalty_shoot_shootout. Goal information and penalty information are extracted from period tables. Goalkeeper records table represents goalkeeper_records table, which contains the details of the game for the goalkeeper. Game statistics table containing the details of the match by player, are represented by game_stats table. Team roster file is about the team information submitted before the beginning of the season. A players_info table containing specific player personal information and a players table containing unique player identifiers were extracted. The Team statistics file was designed with seven tables, each with league-specific information to identify a season's performance. Scoring efficiency statistics are represented in a scoring_efficiency. Statistics under the power play are represented in a power_play. A penalty_killing represents defensive statistics in shorthanded situations and a shorthanded_goal represents offensive statistics in shorthanded situations. A goalkeeping table represents defensive statistics. A penalties table represents penalty statistics. An attendance table represents the number of spectators. The player statistics files show the competitors' respective season records. They were divided into field players and goalkeepers. A playing_statistics and a goalkeeping_statistics were extracted. The league_info, team_info tables representing league information and team information were not available in official result. Their features are used as foreign keys between league and team information when joining the table.

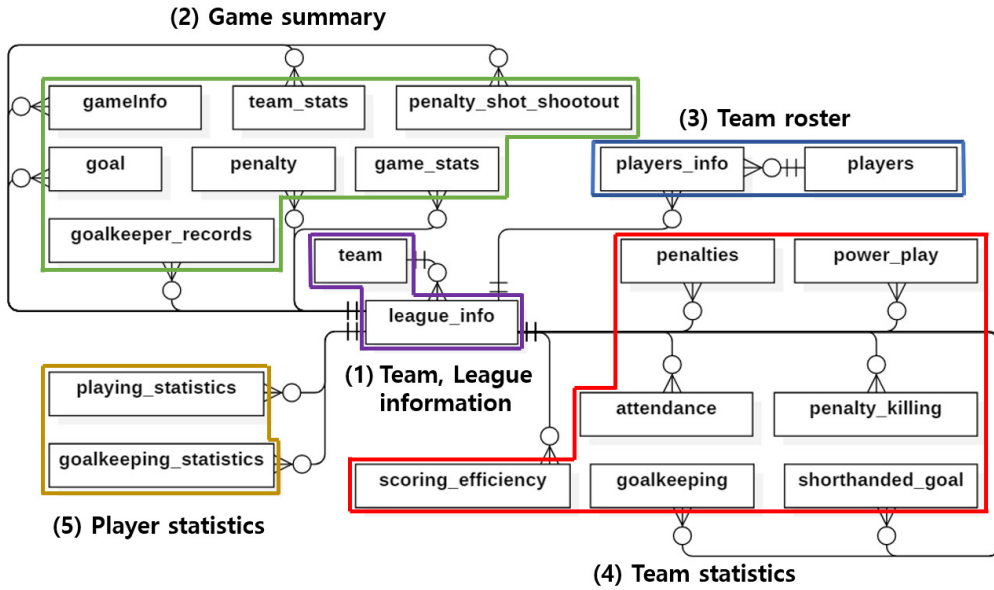


Fig. 2. Ice hockey database schema

scoring_efficiency			power_play			penalty_killing			shorthanded_goals		
PK	tName	char(4)	PK	tName	char(4)	PK	tName	char(4)	PK	tName	char(4)
PK	season	int(4)	PK	season	int(4)	PK	season	int(4)	PK	season	int(4)
PK	sex	char(1)	PK	sex	char(1)	PK	sex	char(1)	PK	sex	char(1)
PK	division	char(8)	PK	division	char(8)	PK	division	char(8)	PK	division	char(8)
	GP	int(4)		GP	int(4)		GP	int(4)		GP	int(4)
	SG	int(4)		ADV	int(4)		DVG	int(4)		SHGF	int(4)
	SSG	int(6)		PPGF	int(4)		PPGA	int(4)		SHGA	int(4)
	SOG	int(6)		PP	text		PK	text		SO	int(4)
	SG	text		TPP	text		TSH	char(8)		ENG	int(4)
				MS	text		MS	text			

goalkeeping			penalties			attendance		
PK	tName	char(4)	PK	tName	char(4)	PK	tName	char(4)
PK	season	int(4)	PK	season	int(4)	PK	season	int(4)
PK	sex	char(1)	PK	sex	char(1)	PK	sex	char(1)
PK	division	char(8)	PK	division	char(8)	PK	division	char(8)
	GP	int(4)		GP	int(4)		spect	int(9)
	GA	int(4)		2	int(4)		GP	int(4)
	SVS1	int(5)		5	int(4)		AVG	int(8)
	SOG	int(5)		10	int(4)			
	SVS2	float		GM	int(4)			
	GAA	float		MP	int(4)			
	SHGA	int(4)		PIM	int(4)			
	PPGA	int(4)		PAVG	text			
	SO	int(4)						
	ENG	int(4)						

Fig. 3. Tables from Team statistics

gameInfo		
PK	order_	int(2)
PK	season	int(4)
PK	sex	char(1)
PK	division	text
	date	text
	overtime	tinyint(1)
	spectators	int(10)
	ref1	text
	ref1Country	text
	ref2	text
	ref2Country	text
	linesmen1	text
	line1Country	text
	linesmen2	text
	line2Country	text

team_stats		
PK	tName	char(4)
PK	season	int(4)
PK	sex	char(1)
PK	division	char(8)
PK	Period	char(16)
PK	order_	int(2)
	Score	int(8)
	SVS	int(8)
	SOG	int(8)
	PIM	int(8)
	TPP	text
	PPG	int(8)
	SHG	int(8)
	win_lose	tinyint(1)
	overtime	tinyint(1)
	date	text
	opp_tName	char(4)

goal		
PK	tName	char(4)
PK	season	int(4)
PK	sex	char(1)
PK	division	char(8)
PK	gid	int(8)
	no	text
	action	char(8)
	period	text
	time	text
	res	varchar(8)
	opp_tName	char(4)
	type	char(10)
	ast1	text
	ast2	text
	flag	tinyint(1)
	date	text
	order_	int(2)
	eng	tinyint(1)

penalty		
PK	tName	char(4)
PK	season	int(4)
PK	sex	char(1)
PK	division	char(8)
PK	pid	int(8)
	no	text
	action	char(8)
	period	text
	time	text
	end_time	text
	opp_tName	char(4)
	type	char(10)
	action_descriptio	text
	date	text
	order_	int(2)

penalty_shot_shootout		
PK	tName	char(4)
PK	season	int(4)
PK	sex	char(1)
PK	division	char(8)
PK	sid	int(8)
	no	int(2)
	time	text
	result	text
	order_	int(2)
	opp_tName	char(4)
	opp_no	text
	action_description	text
	action	char(4)

goalkeeper_records		
PK	tName	char(4)
PK	season	int(4)
PK	sex	char(1)
PK	division	char(8)
PK	No	int(2)
	SVS	int(4)
	SOG	int(4)
	MIP	text
	opp_tName	char(4)
	Name	text

game_stats		
PK	tName	char(4)
PK	season	int(4)
PK	sex	char(1)
PK	division	char(8)
PK	No	int(2)
PK	order_	int(4)
PK	Pos	char(3)
	Name	text
	G	int(4)
	A	int(4)
	P	int(4)
	PIM	int(4)
	FOplus	text
	FOminus	text
	FOnet	text
	FOpercent	text
	sog1	int(4)
	sog2	int(4)
	sog3	int(4)
	sogOT	text
	TS	int(4)
	plusminus	text
	TOT	text
	SHF	text
	AVG	text
	date	text
	opp_tName	char(4)

Fig. 4. Tables from Game summary

players_info		
PK	No	int(2)
PK	season	int(4)
PK	tName	char(4)
PK	sex	char(1)
PK	division	char(8)
	name	text
	year	int(4)
	month	char(4)
	day	int(2)
	pos	char(4)
	weight	int(4)
	height	int(4)
	shoots	char(1)
	team	text

players		
PK	id	int(11)
	name	varchar(255)
	day	int(4)
	month	char(4)
	year	int(4)
	sex	char(1)
	tName	char(4)

Fig. 5. Tables from Team roster

playing_statistics		
PK	tName	char(3)
PK	season	int(4)
PK	sex	char(1)
PK	division	char(8)
PK	no	int(2)
	TSh	text
	TM	text
	SOG	int(4)
	SHG	int(2)
	SGpercent	float
	PTS	int(3)
	PPG	int(2)
	pos	char(2)
	PIM	int(3)
	MdividedbyG	text
	GWG	int(2)
	GP	int(2)
	G	int(2)
	ATdividedbyS	text
	A	int(2)
	net	text

goalkeeping_statistics		
PK	tName	char(3)
PK	season	int(4)
PK	sex	char(1)
PK	division	char(8)
PK	no	int(2)
PK	W	int(2)
	SVSpercent	text
	SVS	int(4)
	SOG	int(4)
	SO	int(2)
	MIPpercent	text
	MIP	text
	L	int(2)
	GPT	int(2)
	GPI	int(2)
	GKD	int(2)
	GAA	float
	GA	int(3)

Fig. 6. Tables from Player statistics

team		
PK	id	int(11)
	tName	char(4)

league_info		
PK	tName	char(4)
PK	sex	char(1)
PK	season	int(4)
PK	division	char(8)

Fig. 7. Team, league information tables

4.2 Raw-data parser development

Parser was developed for the purpose of batch processing to insert the records into the database. Development proceeded with Python 3.6. Four types of files (Game summary, Team roster, Team statistics, Player statistics) were collected out of 22 types of record files. PDF formatted files were converted to CSVs using *tabula-java*, a Java library that extracts tables from PDFs and converts them to CSVs. When importing a CSV file and converting it to pandas DataFrame, different rules were applied as rule base by file and table because different file types had different areas of table recognition. The converted DataFrame was inserted into MySQL using the *sqlalchemy* library.

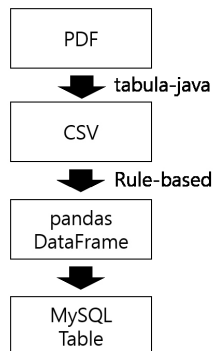


Fig. 8. Ice hockey records parser flowchart

4.3 Developing Ice hockey reference and Validating the data

Unlike the IIHF, the Ice Hockey World Championship statistics site (Ice Hockey Reference) can analyze the trends by providing statistical information collected by teams and competitors for different seasons. In the Ice hockey reference, search conditions are available by season, country, gender and report type. The IIHF provides statistical information about the league for the season in PDF files, but there is a limitation that the competition records cannot be verified in a time series. Meanwhile, the main users of Ice hockey reference are ice hockey coaching staffs and athletes and ice hockey enthusiasts, neither data experts nor ordinary people. Therefore, the site was developed to provide accurate and diverse data while being easy to use and interpret.

Data stored in the database in the form of atomic values through parser is used as raw data in statistical reports for delivery to users. The types of reports to be provided to users were determined

to benchmark NHL STATS and reflect the needs of the sports site. The key functions of statistical services built for data verification consist of player report functions and team report functions. As shown in Table 1, there is a season-by-season statistic for each competitor in the category, and the page associated with the competitor's name can identify the competitor's profile, career record and competition record. In the team category, national records can be checked by season, as shown in Table 2, and in the pages connected, records by year and competition of the respective country can be checked.

Table 1. Report category of Player

No.	Category	Content	Etc
1	Skater Summary	Factors of skater performance	
2	Goals	Scoring situation information	
3	Penalties	Penalty information	
4	Penalties in Detail	Types of penalty, penalty situation information	Field demand
5	Goalie Summary	Factors of goalkeeper performance	
6	Team Roster	Personal information (physical, team info)	
7	1 Period	Factors of periodical performance	Field demand
8	2 Period	Factors of periodical performance	Field demand
9	3 Period	Factors of periodical performance	Field demand
10	Over Time	Factors of periodical performance	Field demand

Table 2. Report category of Team

No.	Category	Content	Etc
1	Team Summary	Factors of team performance	
2	Power Play	Information under Power play	
3	Penalty Killing	Information under Penalty killing	
4	Penalties	Penalty information	
5	Penalties in detail	Specific penalty information	
6	Goalkeeping	Team defensive information	
7	Scoring Efficiency	Scoring efficiency information	
8	1 Period	Factors of periodical performance	Field demand
9	2 Period	Factors of periodical performance	Field demand
10	3 Period	Factors of periodical performance	Field demand
11	Over time	Factors of periodical performance	Field demand

Statistical websites are intended to provide built-in data verification and competition data. Thus, visualization elements (graphs) were added to enable web page users to identify trends in major records. The World Championships information is expected to be updated once a year, with long update cycles. The visualization data of the same league participating countries will be composed

on the main screen, as it is difficult to change the main page screen due to the addition of data. The screen was separated into men/women and three visualizations were made for each. On the five major factors that have a major impact on the game's winning and losing, the top three teams in the league and the top three teams in the league to which Korea belongs are introduced, and in Korea, they provide a graph on the yearly trend (2015, 2017 and 2018). The search UI was referred to in the NHL Statistical Category (<http://www.nhl.com/stats/>).

Table 3. Features for mainpage visualization graphs

Men	Women
SG%	SG%
- Top 3 teams in total	- Top 3 teams in total
- Men's second division	- Women's second division
- Trends by year	- Trends by year
SOG%	SOG%
- Top 3 teams in total	- Top 3 teams in total
- Men's second division	- Women's second division
- Trends by year	- Trends by year
PP%	PP%
- Top 3 teams in total	- Top 3 teams in total
- Men's second division	- Women's second division
- Trends by year	- Trends by year
FO%	FO%
- Top 3 teams in total	- Top 3 teams in total
- Men's second division	- Women's second division
- Trends by year	- Trends by year
PIM	PIM
- Top 3 teams in total	- Top 3 teams in total
- Men's second division	- Women's second division
- Trends by year	- Trends by year

In this study, the raw data quality at the level of element value built with developed parser was verified. As a verification method, statistical and visualization information derived using raw data was verified by ice hockey players and coaches at the sports site. In order to design the interface for data verification, the world championship record including data available to Korean players was selected. The sports statistics site was surveyed, benchmarked, and added the details of the FGI survey with an ice hockey expert. It benchmarked the National Hockey League's statistical search UI (NHL), the world's most popular ice hockey player, and KBL REFERENCE, a domestic basketball community website. The data quality was verified from a sports perspective by reflecting the results of a demand survey conducted on the game records to domestic and foreign ice hockey experts, including the nation's women's national team leaders and players. Table 4 is a comparison of Ice hockey reference and Elite prospects service.

Table 4. Comparison of Elite prospects and Ice hockey reference

Section	Elite prospects	Ice hockey reference
Contest	World championships Asian games Olympic games	World championships
Provided information	30 features (20 of basic information)	66 features
Additions	Player trade information	Types of penalty Periodical statistics Referee information

Elite Prospects (www.eliteprospects.com) is the site that collects competition data for ice hockey competitions, the largest of its kind among ice hockey. At the Korean team level, statistical information can be found in major competitions (World Championships, Asian Games, Olympic Games) and individual information can be found in major competitions. According to the interview results, Korean women’s national team players also check information on the elite Prospects site. It serves as a data hub for ice hockey around the world, including various articles related to ice hockey events, along with statistical information. Elite prospects provide variety information of competitions, but there are fewer types of information available for individual competitions. It is useful to grasp simple trends, but simple information such as goals, assists, and penalty time alone cannot establish the team’s strategy and grasp the players’ performance. The Elite prospects provides a total of 30 kinds of information, but 10 of them are player personal information and 8 are related to the team’s victory or loss, and only 12 are related to the player's and team’s details. On the other hand, Ice hockey reference site provides a total of 66 types of information, enabling specific and practical strategies. Meanwhile, the women’s national team coaches and players wanted to have specific information on the statistics site. The ice hockey reference includes specific types of penalty (boarding, slashing, etc.) and information of judges who gave penalty, and details of matches by career. Also, from the perspective of the national team, ice hockey is usually centered in North America or Europe, so the information of Korea was not detailed, but in the case of ice hockey reference, it is easy to use and develop around Korean players, and it is more useful, including the information that a member of Korean team want to see.

4.4 Core user interviews

In order to meet the practical purpose of this study, future research directions were established through interviews with Korean ice hockey officials. A key part of the interview is extracted and presented in Table 5. Coach Kim Do-yoon of the women’s national team discussed setting up national team data for various age groups, such as the U20 and U18, as well as information from adult players. Coach Kim Do-yoon and women’s national team captain Han Soo-jin hoped to add objectivity to the national team’s participation or annual salary negotiations by assessing performance with accumulated data. Women’s national team members suggested that collecting more diverse leagues

and seasons because world championship is annual event for ice hockey. Team members wanted to compare the results of several competitions. On the other hand, there are two problems that we have found in our research process. First, the key for the competitors provided by the International Ice Hockey Association (IIHF) was not utilized, so the ID for the competitors was re-created via profile, and second, the tabula-java module did not convert the data for 2014 and 2016 into parser. Research is therefore required regarding how to generate competitor IDs and how to extend the functionality of the tabula-java module.

Table 5. Interview with Korean ice hockey officials

Contents	Interviewee
<ul style="list-style-type: none"> • Game records are needed to be collected up to U20 • If competition data is accumulated from youth to adult, Player’s long-term analysis could identify growth trends 	Women’s national team coach Kim Do-yoon
<ul style="list-style-type: none"> • The women’s ice hockey team from Suwon City Hall will play in the middle school league • Records data could be used for salary negotiations 	Suwon Citi Hall coach Kim Do-yoon
<ul style="list-style-type: none"> • After the end of the research, the statistical site should be managed continually 	14 players of Suwon City Hall
<ul style="list-style-type: none"> • This service could be developed to an app to check on smartphone 	14 players of Suwon City Hall
<ul style="list-style-type: none"> • Since the World championships are held once a year, it is likely to log on to the site once and only after the end of the next WC • Homepage usage is likely to be low • Olympic games, friendly matches and Asian league should be added not just World championships • Data surveyor has to observe the competition to obtain detailed record information 	Coaching staffs of Suwon City hall

5. Conclusion

In this study, we analyzed the ice hockey record data to build a database. Ice hockey reference service was established and utilized to verify data deployed by analyzing the original data of the records in PDF format. Record data sources utilized the data stored in the KNSU Datanest data repository. The parser was developed using tablua-java library for batch processing of records. Among the types of records, the game summary, team roster, team statistics, and player statistics files were collected, and tables were extracted from the records. PDF records were converted to text in CSV format. When converting to pandas DataFrame, the rule-based method for each file types was applied to recognize the tables. The converted DataFrame was loaded using the sqlalchemy library. After the above process, out of 22 types of records, four types (Game summary, Team roaster, Team statistics, and Player statistics) including 66830 rowsv of data were deployed. The deployed data verification demonstrated Ice hockey reference service to ice hockey coaching staffs and players

to verify the data. Data verification has found no problems with the deployed data quality and shows a high satisfaction with providing 66 factors against the 30 factors provided by the service previously used. Two issues derived from this study are player identification and the limitation of parser's function depending on the type of record used as raw data. Therefore, further research is needed to overcome the two above problems. Finally, the Ice hockey reference service developed from this study was designed to meet the needs of the sports site and to include specific information at a level not normally utilized. Therefore, it is hoped that it will be actively used in the ice hockey sports field.

References

- Elite Prospects. [n.d.]. Retrieved from <http://www.eliteprospects.com>
- Jin, S., Kim, S. T., & Jang, J. H. (2018). Ice Hockey Database Schema Design: For National Teams's Biomechanical Analysis. *ISBS Proceedings Archive*, 36(1), 610. Retrieved from <https://commons.nmu.edu/isbs/vol36/iss1/138/>
- KBL Reference. [n.d.]. Retrieved from <http://bookyoon.dothome.co.kr/g5/kbl.php>
- Koo, D. H., Panday, S. B., Xu, D. Y., Lee, C. Y., & Kim, H. Y. (2016). Logistic regression of wins and losses in Asia league ice hockey in the 2014-2015 season. *International Journal of Performance Analysis in Sport*, 16(3), 871-880. doi:10.1080/24748668.2016.11868935
- Lee, H. Y. (2010). *A Study on Plans for the Construction of Sports Archive* (Master's thesis). Major of Archival Management, Hanshin University, Osan, Korea. Retrieved from <http://www.riss.kr/link?id=T11984954&outLink=K>
- Memmert, D., & Rein, R. (2018). Match Analysis, Big Data and Tactics: Current Trends in Elite Soccer. *German Journal of Sports Medicine/Deutsche Zeitschrift fur Sportmedizin*, 69(3), 65-72. doi:10.5960/dzsm.2018.322
- Mihindukulasooriya, N., Priyatna, F., & Rico, M. (2018, June). Publishing Tourism Statistics as Linked Data a Case Study of Sri Lanka. In *International Conference on Web Engineering* (pp. 193-201). Springer, Cham. doi:10.1007/978-3-030-03056-8_17
- NHL Stats. [n.d.]. Retrieved from <http://www.nhl.com/stats/>
- Yu, W., Li, Z., Zeng, Q., & Jiang, M. (2019, May). Tablepedia: Automating PDF Table Reading in an Experimental Evidence Exploration and Analytic System. In *The World Wide Web Conference* (pp. 3615-3619). ACM. doi:10.1145/3308558.3314118

[About the authors]

Seung-kyo Jin is a master's degree student in University of Science & Technology at KISTI campus. He is majoring bigdata science. He was contributed to data ecosystem in basketball and ice hockey. He is also data-based columnist for basketball magazine JUMPBALL. The main research field is recommendation systems and deep learning. As a main programmer of Ice hockey Reference,

He has developed records parser to archive ice hockey raw data, designed database schema and statistical website.

Ji-hyun Jang is currently working as a software planner at Argonet. Argonet specializes in providing enterprise content management and information lifecycle support solutions. She holds a bachelor's degree in physics at Ewa Women's University. As a member of the Korea National Sport University, she participated in the "Establishment of Foundation for Interdisciplinary Scientific Research on Winter Sports." She contributed to the paper by identifying the needs of the Korean national ice hockey team and planning statistics sites for the World Ice Hockey Championships by referring to domestic and international sports statistics sites.

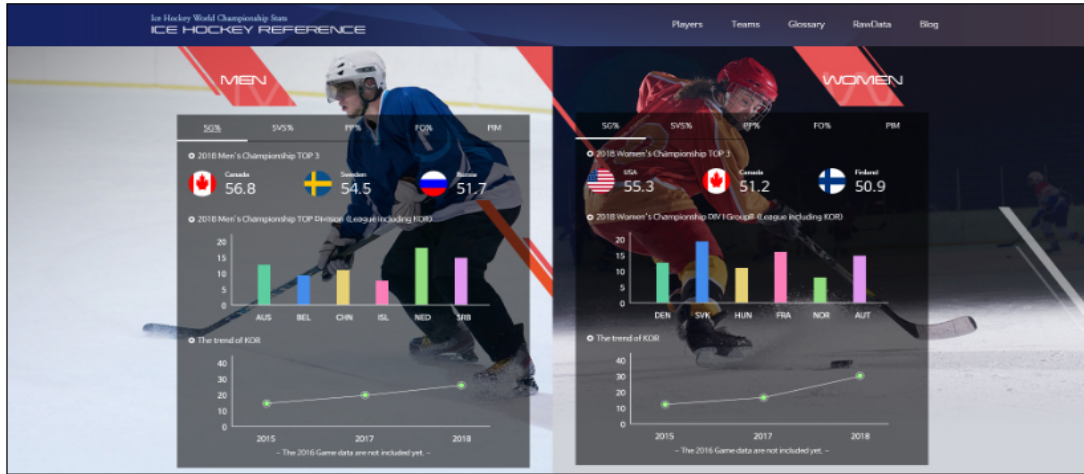
Hye-young Kim was graduated from Ewha Womans University in 1983 with a doctorate in applied physics. Since 1995, she has been working at Korea National Sport University and teaching sports physics and natural science as a basic liberal arts education. And she is interested in R&D policy and convergence research necessary to vitalize sports industry based on sports science and technology by linking science and technology with sports field, including sports science research to improve performance of elite athletes. These days, she is working on sports data research, and serves as vice president of The Korean Association of General Education and vice president of the Korea Ice Hockey Association.

Sun-tae Kim is an assistant professor in the Library and Information Science Department of Jeonbuk National University. He received his PhD Degree in Library and Information Science from the Jeonbuk National University. Before his current appointment, He worked for Korea Institute of Science and Technology Information (KISTI) as Principle Researcher and he worked as a computer program developer at Linksoft which developed the NDSL and NOS. His research interests include research data management, research data platform, research data sharing, semantic web, metadata. Dr Suntae Kim can be contacted at: kim.suntae@jbnu.ac.kr.

Acknowledgement

This research was supported by the Sports Science Convergence Technology Development Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2014M3C1B1034027).

[Appendix] Ice Hockey Reference



Appendix 1. Ice hockey reference main page

The screenshot shows the 'Teams' page for the 2018 season. It features a table with the following columns: Team, Season, Ice, Division, GP, W, OTW, OTL, L, PTS, SOG, FO+, FO-, FO+, FO%, GDF, GF, GA, GF/GP, GA/GP, P/N, P/P, SOG, P/M, Pen. The table lists statistics for teams KOR, USA, LAT, KAZ, POL, and CHN.

Team	Season	Ice	Division	GP	W	OTW	OTL	L	PTS	SOG	FO+	FO-	FO+	FO%	GDF	GF	GA	GF/GP	GA/GP	P/N	P/P	SOG	P/M	Pen
KOR	2018	F	IB	5	3	1	0	1	14	156	192	133	79	0.63	12	28	8	4.00	1.80	95.45	26.67	12.82	42	21
USA	2018	F	IB	5	4	0	0	1	12	128	115	188	-73	0.38	8	14	6	2.80	1.20	33.58	16.00	10.94	42	21
LAT	2018	F	IB	5	3	0	0	2	9	161	150	142	8	0.51	-5	7	12	1.40	2.40	87.50	19.05	6.83	19	18
KAZ	2018	F	IB	5	2	0	1	2	7	144	131	165	-34	0.44	-1	9	10	1.80	2.20	66.67	5.56	6.25	24	12
POL	2018	F	IB	5	0	0	0	5	0	112	156	132	24	0.54	-14	7	21	1.40	4.20	77.78	6.67	6.25	40	20
CHN	2018	F	IB	5	2	0	0	3	6	140	140	144	-4	0.49	0	7	7	1.40	1.40	80.00	0.00	4.96	32	16

Appendix 2. Ice hockey reference teams page

The screenshot shows the 'Players' page for the 2018 season. It features a table with the following columns: No, Player, Season, Team, Division, POS, Shots, S, GF, G, A, P, P/GP, P/M, Pen, FO+, FO-, FO+, SOG, SOG, +/-.

No	Player	Season	Team	Division	POS	Shots	S	GF	G	A	P	P/GP	P/M	Pen	FO+	FO-	FO+	SOG	SOG	+/-	
1	MALCOLM Gault	2018	USA	IB	CF	L	F	5	0	0	0	0.00	0	0	0	0	0	0	0	0.00	0
2	MATTW Kaula	2018	USA	IB	D	L	F	5	2	1	3	2.00	2	1	0	2	-2	22	0.66	3	
3	BECCA Valbruna	2018	USA	IB	D	L	F	5	0	0	8	0.00	4	2	0	0	0	1	0.00	0	
4	SALTTA Cantu	2018	USA	IB	F	L	F	5	2	2	4	2.40	4	2	46	46	3	0	0.25	5	
5	de RECK Miro	2018	USA	IB	D	L	F	5	0	0	8	0.00	0	0	0	0	0	0	0.00	0	
6	GALTERI Federico	2018	USA	IB	D	R	F	5	1	1	3	1.20	0	0	7	-7	7	0.04	-1		
7	de ROCCO Usher	2018	USA	IB	D	L	F	5	2	1	3	2.00	0	0	0	0	0	7	0.25	3	
8	CALMO Anna	2018	USA	IB	F	L	F	5	0	0	8	0.00	2	1	5	23	-16	0	0.00	-1	
9	de la FORREST de ERIGNON Anna	2018	USA	IB	F	L	F	5	3	2	5	2.40	4	2	3	12	-9	16	0.16	7	
10	MURARO Aude	2018	USA	IB	F	L	F	5	0	0	8	0.00	0	0	1	0	1	5	0.00	0	
11	ROCCOJA Rebecca	2018	USA	IB	F	L	F	5	0	1	1	0.00	4	2	0	0	0	5	0.00	-2	
12	BATTISTI Georgia	2018	USA	IB	F	L	F	5	0	0	8	0.00	0	0	0	0	0	2	0.00	-1	
13	CARONARD SARA	2018	USA	IB	F	L	F	5	1	0	1	1.00	2	1	27	28	-11	0	0.12	-2	
14	MAGRIBI Muelita	2018	USA	IB	D	R	F	5	0	0	8	0.00	0	0	0	0	0	1	0.00	-1	

Appendix 3. Ice hockey reference players page

Ice Hockey World Championship Sites
 ICE HOCKEY REFERENCE

Players Teams Glossary RawData Blog

CHOI Jiyeon

10 H F 1.58 H 61 W

Birth: AUG 21, 1998 Shoots / Gatches: L Club: Ice Beet

Career Game Logs

Season	Team	Division	GP	G	A	P	+/-	PIM	PPG	SHG	SHF	GWG	OTG	S	SA	FO%
2015	KOR	SA	2	2	2	0	0	4	0	1	0	1	0	20	6.0	3.0
2016	KOR	SA	5	1	2	3	5	8	0	0	1	0	0	10	6.0	3.0
2017	KOR	SA	5	1	2	3	1	4	1	2	0	1	0	14	6.0	-

Appendix 4. Ice hockey reference players career