Received: 8 February 2021 | Accepted: 5 May 2021 | Published online: 28 June 2021

DOI: 10.1002/its2.76

International Turfgrass Society Research Journal

TECHNOLOGY INTEGRATION IN TURFGRASS MANAGEMENT

NTEP-DB 1.0: A relational database for the national turfgrass evaluation program

Yiqun Xie[1,2,3] | **Majid Farhadloo**[1] | **Ning Guo**[1] | **Shashi Shekhar**[1] | **Eric Watkins**[4] | **Len Kne**[5] | **Han Bao**[5] | **Aaron J. Patton**[6] | **Kevin Morris**[7]

[1] Dep. of Computer Science and Engineering, Univ. of Minnesota, Minneapolis, MN 55455, USA

[2] Dep. of Geographical Sciences, Univ. of Maryland, College Park, MD 20740, USA

[3] Center for Geospatial Information Science, Univ. of Maryland, College Park, MD 20740, USA

[4] Dep. of Horticultural Science, Univ. of Minnesota, St. Paul, MN 55108, USA

[5] U-Spatial, Univ. of Minnesota, Minneapolis, MN 55455, USA

[6] Dep. of Horticulture and Landscape Architecture, Purdue Univ., West Lafayette, IN 47907, USA

[7] National Turfgrass Evaluation Program, Beltsville, MD 20705, USA

Correspondence
Yiqun Xie, Dep. of Computer Science and Engineering, Univ. of Minnesota, Minneapolis, MN, 55455 USA.
Email: xie@umd.edu

Abstract

Given field experiment data collected by the National Turfgrass Evaluation Program (NTEP), we aim to design and create a relational database to store the data and support efficient queries. As one of the most widely-known turfgrass research programs in the world, NTEP has generated large volumes of data on turfgrass cultivars and experimental germplasm since the early 1980s, providing invaluable information for a variety of user groups (e.g., homeowners, seed companies, golf course managers, retailers, turfgrass researchers) to select cultivars that best fit their needs (e.g., winter survival, pest tolerance, turf quality). The datasets have historically been stored in large sets of text files and spreadsheets. Currently, NTEP data are delivered to users through a website (www.ntep.org) as summary reports and it can be extremely tedious (e.g., hundreds of clicks, data merging, jargon) to perform a simple query (e.g., best cultivar selection with typical conditions). This significantly limits the use of NTEP data and hides its value from the public. To address these limitations, we carried out an interdisciplinary effort with horticulture and computer science researchers to design and create the first NTEP database – NTEP-DB 1.0 – to reduce the manual efforts and expert knowledge currently required to extract meaningful information from the data. Experiments confirm that the query results are correct, and that the database can greatly reduce manual efforts. Anticipating next-generation advances, we also recommend incorporating spatial data types and analytical techniques into future designs of the database.

1 | INTRODUCTION

The National Turfgrass Evaluation Program (NTEP), one of the most widely-known turfgrass research programs in the world, has coordinated multi-location turfgrass cultivar evaluation trials since 1981 (Morris & Shearman, 2000; NTEP, 2020a). Each year, NTEP sponsors the establishment of new turfgrass cultivar evaluation trials by requesting sponsors (seed companies or public turfgrass breeding programs) to submit experimental varieties or named cultivars. More than twenty turfgrass species, encompassing over 2,500 experimental selections and cultivars have been tested since the program's establishment. Submitted entries are then sent to multiple testing locations in the U.S. and Canada, most of which are associated with University-based turfgrass research programs. Collaborators at these institutions establish and manage the trials, collecting performance data, which they submit to NTEP at the end of each growing season. Data from

these multi-year trials provide very rich information that are important for decision-making in a variety of domain applications (Table 1). Homeowners, for example, can search for information on a cultivar that best suits their needs/preferences based on a variety of conditions, including local weather (e.g., winter hardy species have better chance of survival in cold climates), maintenance requirement, traffic tolerance, density, color, pest tolerance, etc.

For about 40 years, data collected by NTEP have been stored in large sets of text files and spreadsheets. Currently, these data are delivered to users through a website (www.ntep.org) as summary reports. Figure 1a and b show the general look of the website and an example summary report on quality ratings of fine fescue cultivars in five different US locations in 2016.

Although the summary reports provide consumers (e.g., homeowners, golf course superintendents) with meaningful information (e.g., color, disease resistance) to select a desired adaptation of a cultivar, the data are currently underutilized because the website's flat-table representation is highly intimidating to the average homeowner and many turfgrass professionals. For example, to compare a set of turfgrasses at a certain location, consumers need to go through multiple levels of web pages from the current NTEP website, manually locate relevant table names and columns, record the values and finally make comparisons. Figure 2 traces the manual work needed to complete the following simple query: List the names of cultivars with an average quality rating of at least 6 (scale is 1–9 where 9 is outstanding or ideal turf, 6 is considered acceptable, and 1 is poorest or dead) from 2004 to 2014 in Minnesota.

As shown in Figure 2, the number of clicks needed to retrieve quality ratings for one type of grass in one state in a single year is 4. Given that there are 15 types (from the NTEP website) and 10 years, as specified by the query, about 600 manual clicks are needed to collect the needed quality rating information. Furthermore, this does not include the manual work needed to search and filter out cultivar names with a quality rating higher than or equal to 6 from the data in the 600 summary reports. Clearly, this process is extremely tedious, especially since most users have become accustomed to retrieving other types of information based on simple queries (e.g., hotel prices, local restaurants). This tedious process significantly discourages users from exploring a larger volume of the data provided by NTEP, which defeats a major purpose of the program.

A recent survey of NTEP users found that there was a disconnect between the information NTEP was providing and the knowledge consumers had about grass seed purchasing (Yue et al., 2019). Most survey respondents were familiar with NTEP and the data they provided; however, the number of respondents who actually visited the NTEP website to obtain data was very low. NTEP was providing a lot of data, but the data were not being used by consumers. Respondents desired a better data output format for NTEP data. Similarly, an earlier survey found that consumers have very limited knowledge about purchasing grass seed for their lawns (Yue et al., 2017). It is clear that NTEP's current data practices do not align with the needs of consumers. To maximize the potential of NTEP data, there is an urgent need to have a complete redesign of its data storage and management, as well as a new way to deliver information to users.

Core Ideas

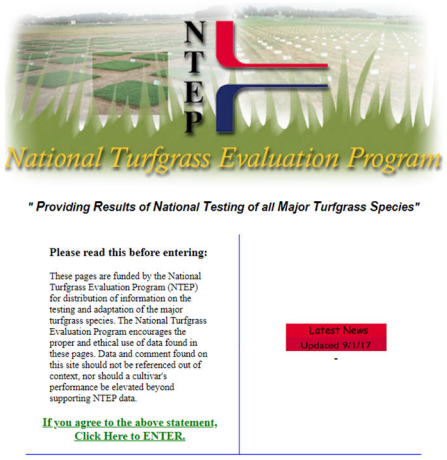
- Created the first database for the National Turfgrass Evaluation Program (NTEP)
- Significantly reduced the effort of information extraction for broad NTEP users
- Validated the database via correctness and flexibility tests
- Provided recommendations for the next generation of NTEP database

Our interdisciplinary effort, which included team members from computer and horticultural sciences explores the opportunity to move NTEP data to a modern relational database and use the data in new and innovative ways that can effectively and efficiently answer a wide range of queries from users. There are two major reasons that we choose a relational database. First, it provides a robust and efficient way of data management and querying with major benefits such as (1) high data quality with constraints (e.g., integrity constraint); (2) improved data safety with automatic backups; and (3) improved data security with access control. With a relational database, it also becomes easier and more efficient to collect and share data (e.g., uploading data directly from smartphones in the field) and generate outputs in various desired formats (e.g., spreadsheets, maps, software-specific formats). Second, relational databases have long been tested in real-world applications, and they are widely used by many large organizations due to the data quality benefits. Some examples include agriculture (e.g., USDA, Monsanto), healthcare (e.g., National Center for Biotechnology Information), airlines (e.g., Delta, American Airlines), banks (e.g., Bank of America, US Bank), retailers (e.g., Target, Walmart), and most universities.

There are three major challenges in designing the database. First, an appropriate database design needs knowledge of NTEP user preferences (e.g., frequently used queries, output formats). Second, data collections are not always static; although NTEP has collected a large set of data, there are new data types that should be collected and incorporated into the database (e.g., new experiments, new species, newly tested properties, new spatial information). For instance, inclusion

TABLE 1 Examples of use cases of NTEP data

User Group	Use Cases
General public (e.g., homeowners)	Select grasses that work well in their home lawn.
Retailers	Determine which cultivars to stock for sale based on their location.
Turfgrass managers (e.g., golf course managers)	Identify turfgrass cultivars with best potential at their site.
Seed companies	Promote the production and use of better cultivars.
Researchers (e.g., horticultural science, computer science)	Horticultural science: summarize data and perform analysis to understand turfgrass performance and possible targets for improvement through plant breeding; Computer science: explore advanced data science techniques (e.g., spatial data mining, machine learning) to detect patterns (e.g., hotspots) and make predictions.
Extension specialists and consultants	Summarize data and perform analysis to understand turfgrass performance and make recommendations to clients.



(a) Main page of the NTEP website

TABLE 2A. TURFGRASS QUALITY RATINGS OF FINELEAF FESCUE CULTIVARS 1/
GROWN IN LOCATIONS PERFORMANCE INDEX (LPI) GROUP 2 */
2015 DATA
TURFGRASS QUALITY RATINGS 1-9; 9=IDEAL TURF 2/

NAME	MI1	MI2	MN2	ND1	WA1	MEAN
BEUDIN	6.7	6.9	6.2	6.4	6.0	6.4
BAR FRT 5002	6.4	7.3	6.1	5.7	5.6	6.2
GLADIATOR (TH456)	6.4	6.1	5.8	6.6	5.8	6.1
RESOLUTE (7H7)	6.3	6.4	5.9	6.3	5.6	6.1
DLF-FRC 3338	6.2	6.9	6.0	5.7	5.6	6.1
PPG-FRC 113	6.1	7.0	6.0	5.6	5.5	6.1
RAD-FC44	6.2	6.7	5.9	5.8	5.6	6.0
RADAR	6.1	6.7	5.9	5.8	5.6	6.0
BAR VV-VP3-CT	6.2	6.7	5.9	5.8	5.5	6.0
PPG-FRT 101	6.1	6.7	5.9	5.8	5.6	6.0
PPG-FRC 114	6.1	6.9	5.9	5.6	5.5	6.0

(b) Example of a summary report

FIGURE 1 NTEP website and an example summary report

of spatial information for every tested plot of grass (GPS locations; surveyed X,Y coordinates of plot vertices) would allow for richer spatial analysis in the future. Lastly, NTEP users may not be familiar with query languages such as SQL, and some users may not have the motivation to learn it. For example, while NTEP managers and researchers may be willing to learn SQL for better data management and analyses, consumers (e.g., homeowners) are unlikely to learn a query language just to select an ideal turfgrass for their lawn.

To address these challenges, we first gained understanding about NTEP users from an existing survey work, which covered over 400 participants (Yue et al., 2019), and then carried out the following projects. We gather NTEP data, design (i.e., conceptual, logical and physical) and implement the first NTEP database, namely NTEP-DB-1.0, based on many rounds of interdisciplinary and multi-sectoral discussions among researchers from computer science, horticultural science and geographic information science, as well as users. NTEP-DB 1.0 converts the NTEP webpage-based summary reports to a relational database, and allows researchers and other turf professionals to quickly access desired information using queries.

To validate the database, we carried out two sets of experiments. First, for a sanity check, we perform a set of queries to reproduce some existing summary reports in the current NTEP website and compare them to see if the numbers/statistics are the same. Second, for a usability and flexibility check, we have real users (e.g., consumers, turfgrass researchers) provide a wide range of types of queries to test if they can be answered by the database. Experiment results confirmed both the correctness and flexibility of the database.

Finally, to better meet the needs of non-technical users, we recommend a new web-based interface (with a prototype example), which uses the database as a core and allows users (e.g., consumers, retailers) to extract the information they need without knowing any jargons or writing any query language such as SQL. In addition, we make two more recommendations for the next generation of the NTEP database, i.e., real-time data updates and a spatial database (e.g., spatial data structures, spatial statistics, and data mining techniques),

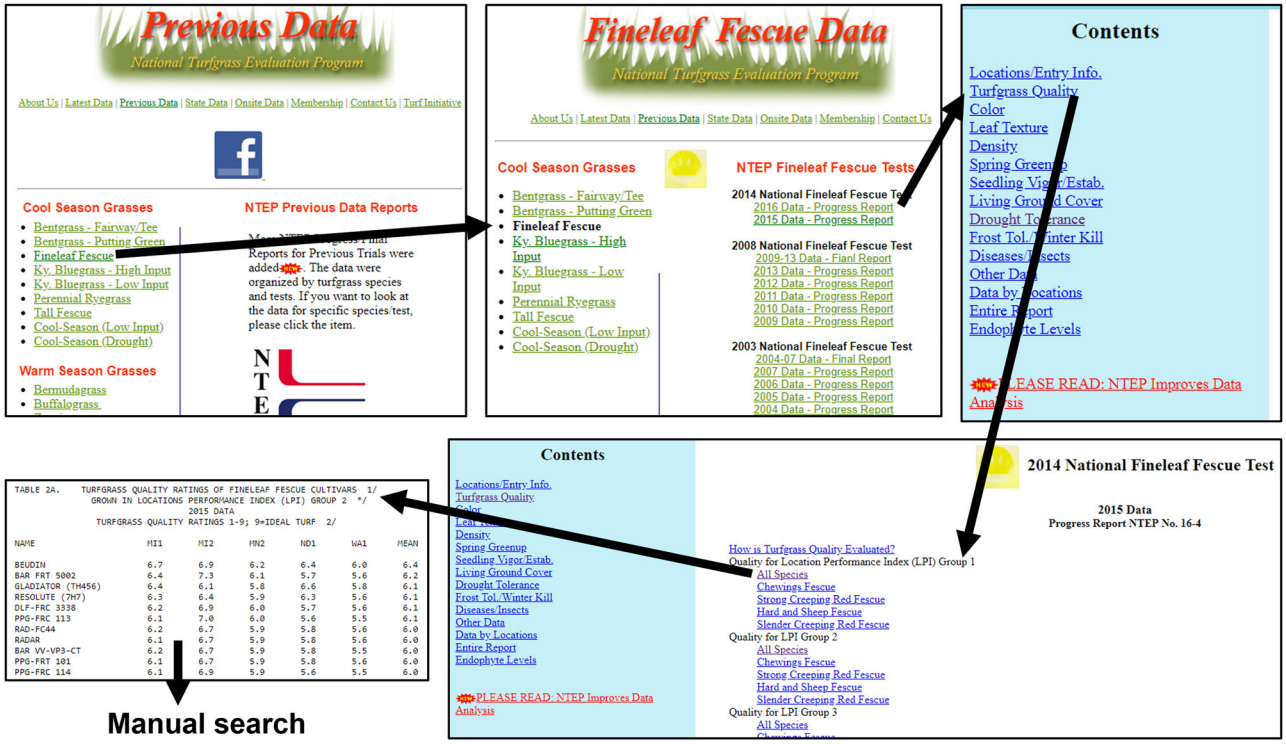


FIGURE 2 An example trace of manual clicks needed to retrieve quality ratings of one species in one state in a single year

which can further improve data quality, reduce management effort, and help plant breeders and researchers to better understand the spatial autocorrelation and variability in turfgrass performance.

1.1 | Scope

This work focuses on designing and creating the first version of the NTEP database, which is a major milestone for both NTEP and its user communities (e.g., consumers, Extension specialists, turfgrass managers, seed companies, researchers). This paper aims to introduce this new database to the turfgrass community. Advancing database research in computer science is outside the scope of the present study and will be explored in future work based on the domain science needs of the turfgrass community (e.g., spatiotemporal analysis, interactions between genetics and spatial environment). This work also does not consider data analytics methods that can be further applied on the output of a database query.

2 | MATERIALS AND METHODS

In this section, we describe the steps to convert the NTEP webpage-based summary reports to a relational database – NTEP-DB 1.0 – through three levels of database design: conceptual, logical and physical.

2.1 | A first step: From NTEP data to a single wide table

To understand the actual NTEP data, we first gathered all data from NTEP, which are stored on a single spreadsheet with all the attributes. These data were then inserted, in their original form, into a database as a single wide table where each column represents a single attribute. Figure 3 shows a list of example attributes in this wide table.

While the single wide table can already be used to provide answers to some simple queries in a much easier manner compared to manual search-and-filtering through NTEP webpages (Figure 2), it suffers from main issues related to data quality and query flexibility (Navathe & Elmasri, 2001; Shekhar & Chawla, 2003).

2.1.1 | High risk of data inconsistency

A typical data quality issue that arises when all information is stored in a single table is **data redundancy**, i.e., attributes are duplicated because they need to be stored together with other attributes upon which they do not depend (Navathe & Elmasri, 2001; Shekhar & Chawla, 2003). In the context of NTEP data, for example, "cultivar name" and "site ID" should not depend on an environment attribute such as "shade condition". However, such dependency might be implicitly assumed when they

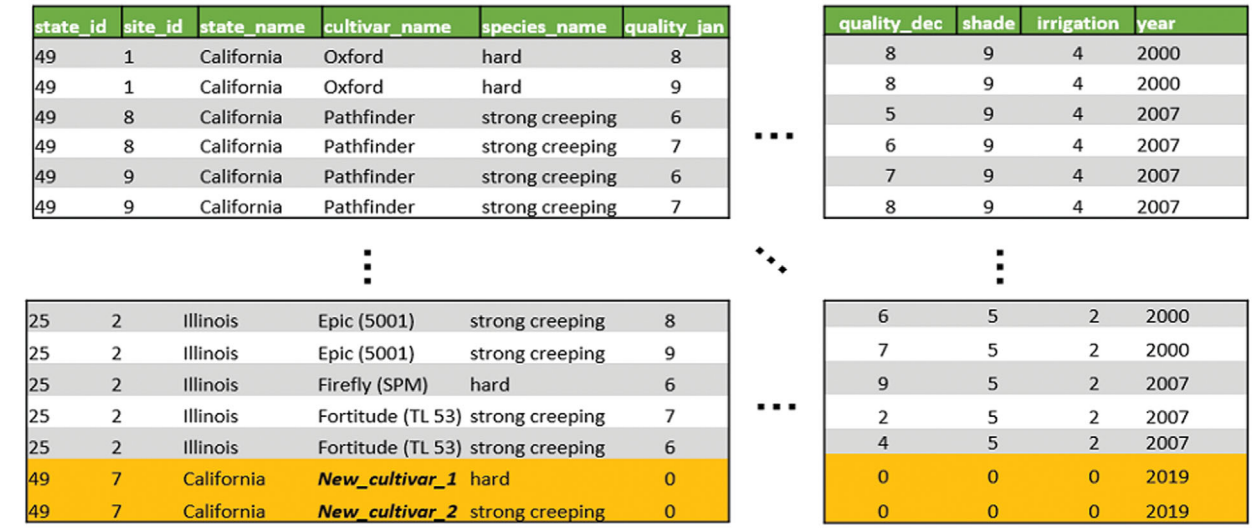
FIGURE 3 Example attributes in the single wide table

are stored together in a single table. This problem can cause three types of data inconsistency issues for common database management operations:

- *Data insertion*: (1) If the wide table allows adding rows with some attributes being left empty (e.g., filled by 0 or empty strings), insertions can easily distort summary statistics about these attributes. In Figure 4a, information about two new cultivars is added to the wide table prior to the completion of their field experiments. Since their "quality_rating" values are not available, the corresponding columns are filled with 0 by default. Consequently, any summary statistic involving these "0" values will lead to a spurious estimate. Figure 4b shows another example where two records of a single cultivar ("Pathfinder" highlighted in blue) are added with available quality ratings. However, since the "cultivar_name" column allows duplicates, which is necessary for the wide table, the database administrator did not realize the name is a duplicate of an existing name shown in the top part of Figure 4b thereby treating two different cultivars as a single cultivar, which would lead to distorted summary statistics. (2) If the wide table does not allow new rows with empty values, the design prevents a database administrator from inserting information about a new cultivar. This is inconvenient since any new information has to be suspended until all information across over a hundred attributes is available. This may potentially lead to information loss in the database.
- *Data update*: When a site ID needs to be changed (e.g., due to relocation or split), it has to be changed in many places in the table due to the fact that it is stored together with 100+ other attributes whose values may have a large number of different combinations (i.e., rows) in the table. If a database administrator misses any of these duplicated places, the same site in the real-world may correspond to multiple IDs in the database without anyone knowing. Such errors may not even be recoverable and can significantly reduce data quality.
- *Data deletion*: When a database administrator would like to delete undesired shade conditions (e.g., −1), this may lead to unexpected deletion of cultivar names and site IDs that were only associated with such shade conditions in the data table. As a result, these cultivar names and site IDs will completely disappear and become unsearchable.

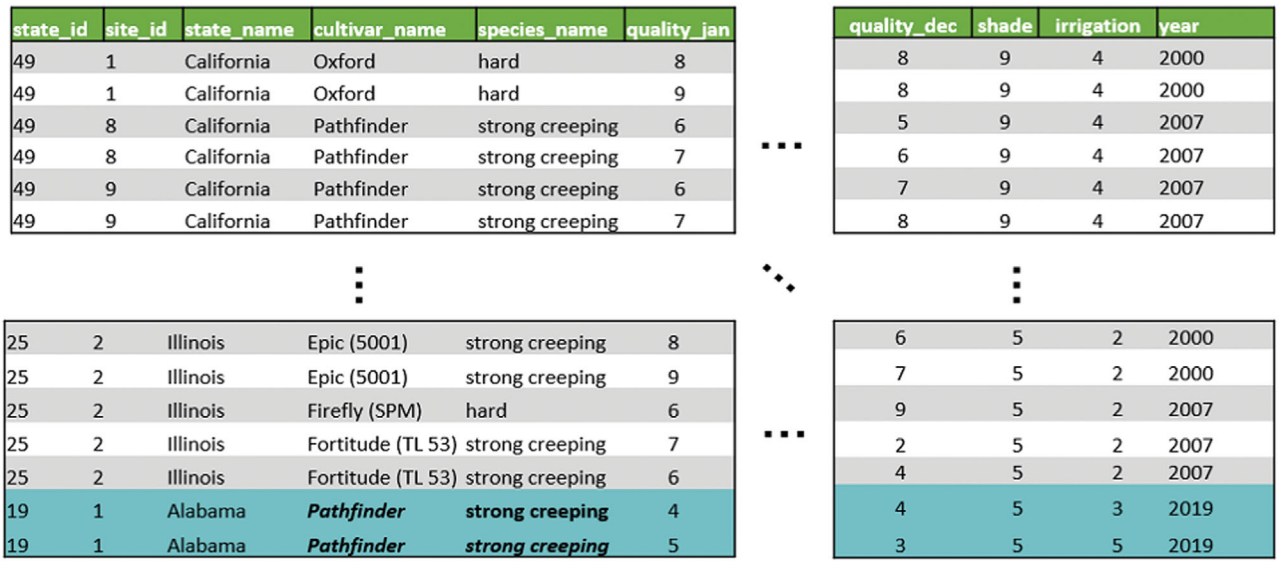
2.1.2 | Limited query flexibility

The column design in the single wide table, which is directly inherited from the original NTEP data (e.g., spreadsheets, text files), also limits the set of queries that can be easily answered using database query language SQL. For example, the quality ratings for different months are stored as different columns in the wide table (e.g., "quality_rating_january", "quality_rating_february"). This design discourages many frequently used query types that require summary statistics across several months because aggregation functions (e.g., "AVERAGE, "SUM", "STD") and ranking functions (e.g., "ORDER BY") in SQL typically operate on a single column and do not expand well across columns. The following shows a list of example common queries that are limited by the wide table design:



state_id	site_id	state_name	cultivar_name	species_name	quality_jan	quality_dec	shade	irrigation	year
49	1	California	Oxford	hard	8	8	9	4	2000
49	1	California	Oxford	hard	9	8	9	4	2000
49	8	California	Pathfinder	strong creeping	6	5	9	4	2007
49	8	California	Pathfinder	strong creeping	7	6	9	4	2007
49	9	California	Pathfinder	strong creeping	6	7	9	4	2007
49	9	California	Pathfinder	strong creeping	7	8	9	4	2007
25	2	Illinois	Epic (5001)	strong creeping	8	6	5	2	2000
25	2	Illinois	Epic (5001)	strong creeping	9	7	5	2	2000
25	2	Illinois	Firefly (SPM)	hard	6	9	5	2	2007
25	2	Illinois	Fortitude (TL 53)	strong creeping	7	2	5	2	2007
25	2	Illinois	Fortitude (TL 53)	strong creeping	6	4	5	2	2007
49	7	California	***New_cultivar_1***	hard	0	0	0	0	2019
49	7	California	***New_cultivar_2***	strong creeping	0	0	0	0	2019

(a) Insertion of two new cultivars with missing values



state_id	site_id	state_name	cultivar_name	species_name	quality_jan	quality_dec	shade	irrigation	year
49	1	California	Oxford	hard	8	8	9	4	2000
49	1	California	Oxford	hard	9	8	9	4	2000
49	8	California	Pathfinder	strong creeping	6	5	9	4	2007
49	8	California	Pathfinder	strong creeping	7	6	9	4	2007
49	9	California	Pathfinder	strong creeping	6	7	9	4	2007
49	9	California	Pathfinder	strong creeping	7	8	9	4	2007
25	2	Illinois	Epic (5001)	strong creeping	8	6	5	2	2000
25	2	Illinois	Epic (5001)	strong creeping	9	7	5	2	2000
25	2	Illinois	Firefly (SPM)	hard	6	9	5	2	2007
25	2	Illinois	Fortitude (TL 53)	strong creeping	7	2	5	2	2007
25	2	Illinois	Fortitude (TL 53)	strong creeping	6	4	5	2	2007
19	1	Alabama	***Pathfinder***	**strong creeping**	4	4	5	3	2019
19	1	Alabama	***Pathfinder***	***strong creeping***	5	3	5	5	2019

(b) Insertion of two new cultivars with name duplicating existing ones

FIGURE 4 Data consistency problems in the context of NTEP data in a flat table

- Which cultivar has the highest average quality rating in summer (i.e., June, July, August)?
- List the average color rating for fine fescue cultivars in 2016.
- Which disease (also stored as different columns) affects cultivars of "slender creeping red fescue" the most?

2.2 | A conceptual level design with an entity-relationship diagram

The key issue of the single wide table design is data redundancy, which as we introduced, leads to high risk of data inconsistency and limited query flexibility. In database design, the process to remove such redundancy is called "normalization", which splits a single table into a set of narrower tables with fewer columns (Navathe & Elmasri, 2001; Shekhar & Chawla, 2003; Mannino, 2005; Garcia-Molina et al., 2000). Each table then corresponds to a single entity in the targeted domain (e.g., a "cultivar" entity in the case of turfgrass) and there are links between tables to connect them (e.g., an experiment trial "evaluates" a set of cultivars).

Such a design typically requires a combination of database expertise as well as the knowledge about the application domain (i.e., turfgrass) to satisfy two design requirements:

- **Satisfaction of normal forms**: Normal forms are database normalization criteria used to explicitly control the level

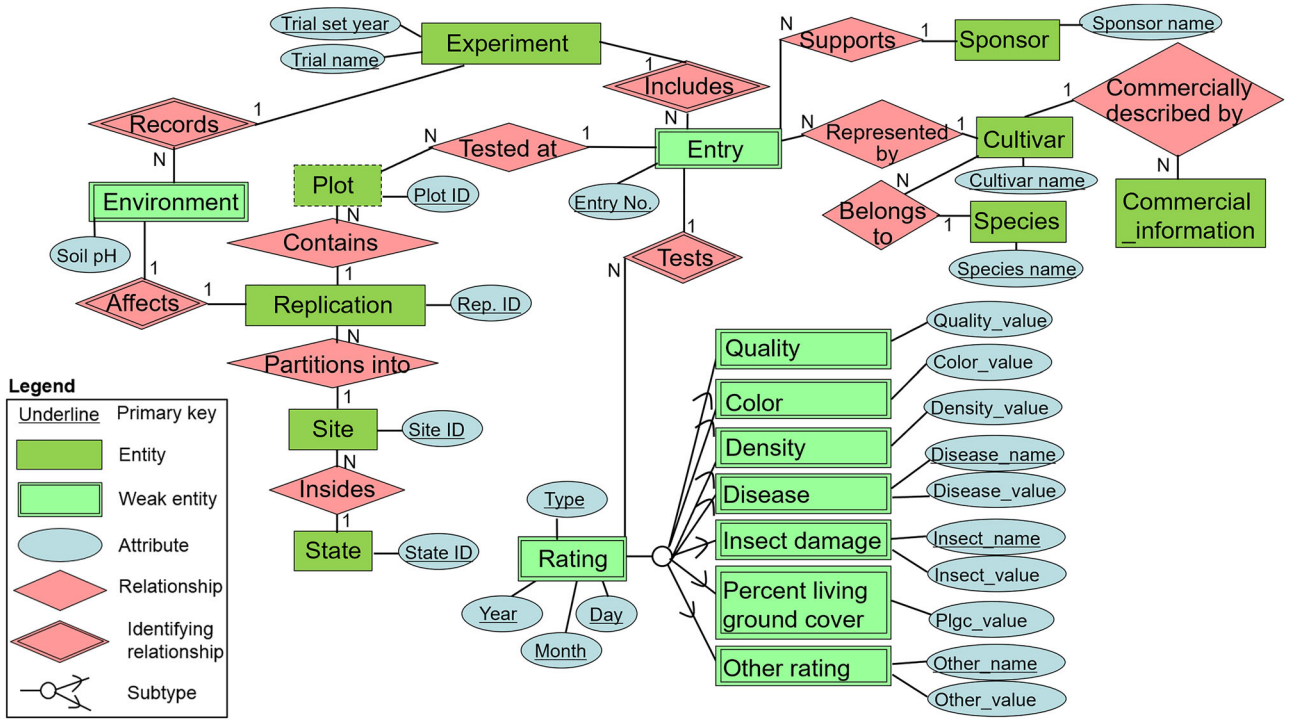


FIGURE 5 Comprehensive Entity-Relationship diagram of NTEP database

of data redundancy. We implemented the most commonly used and recommended "3rd normal form" to define the criteria. Denote "**primary key**" as a minimal set of columns that can uniquely identify each row in a table. The norm requires that all columns in the table depend on the primary key, the whole primary key, and nothing but the primary key. The wide table violates this condition. For example, a quality rating depends on a combination of spatial, temporal, environmental and cultivar related attributes (columns), whereas a species name can be uniquely identified by a cultivar name (i.e., a subset of the previous set, indicating that the subset must not be the whole primary key). This difference in dependencies makes it unable to satisfy the norm.

- **Domain interpretability**: The design also needs to be easily interpretable to domain experts (i.e., turfgrass professionals) so they can easily verify the correctness of the entities as well as their links, and check the dependencies. Domain interpretability also helps users to identify new opportunities to improve data collection and experiment design (e.g., adding new environmental attributes or new spatial information as discussed later). This in turn helps make the database design more flexible for potential future extensions.

We use an entity-relationship diagram (ERD) to meet these requirements (Li & Chen, 2009; Navathe & Elmasri, 2001). In the following, we first introduce the key concepts in ERD (i.e., entity, relationship and cardinality constraint), and then show the ERD design for NTEP-DB-1.0.

An **entity** (green rectangles in Figure 5) in ERD refers to a real-world object or concept (e.g., a cultivar or an experiment trial in NTEP). Each entity has a set of attributes. As we introduced above, the minimal set of attributes that can be used to uniquely identify an instance of an entity is called the primary key of the entity (e.g., "cultivar_name" can be the primary key for the entity "cultivar"). A **relationship** (pink diamonds in Figure 5) is the link between two entities. For example, a cultivar "belongs to" a species, a site is inside a state, and a cultivar in an experiment is supported by a sponsor. Finally, there are **cardinality constraints** (numbers in Figure 5) that restrict the number of entity instances participating in a relationship. There are three types of cardinality constraints: (1) One-to-one: Each instance of entity A can only connect to one instance of entity B and vice versa (e.g., a cultivar is represented by one and only one entry number in an experiment); (2) One-to-many: Each instance of entity A can connect to many instances of entity B but each instance of B can only relate to a single instance of A (e.g., a site can only be inside one state but a state can have multiple sites); and (3) Many-to-many: Instances of both entities A and B can connect to many instances of the other.

Figure 5 shows the entity-relationship diagram we designed for NTEP-DB-1.0, and the explanations of entities and relationships are provided in Tables 2 and 3, respectively. The blue ellipses represent the attributes of each entity. Since there are over a hundred attributes which cannot fit in the figure, we only listed the primary key (underlined) and a few example attributes for each entity (primary keys of weak entities and

TABLE 2 Entity names and descriptions in the NTEP database

Entity Name	Description
State	A state in the U.S.
Site	A turfgrass trial site location in a state
Replication	A partition of a site so that each entry has multiple instances across these partitions
Entry	The ID of a cultivar in a species trial
Cultivar	The most fine-scaled classification of grasses in the database
Species	The grass species, each contains a distinct set of cultivars
Commercial_information	The commercial information for a cultivar
Sponsor	An organization that pays the testing fees for specific cultivars
Experiment	A 5-year trial, determined by a trial name and trial set year
Environment	The growing environment or conditions of the grass (e.g., soil pH)
Rating	The rating for a set of grass attributes
Color	The color rating of an experimental unit
Quality	The quality rating of an experimental unit
Density	The density rating of an experimental unit
Disease	The disease (multiple) rating of an experimental unit
Insect_damage	The insect damage (multiple) rating of an experimental unit
Percent_living_ground_cover	The percent of living ground cover of an experimental unit
Other_attributes	The rating of other attributes (e.g., green up) of an experimental unit

TABLE 3 Relationship descriptions in NTEP data between pairs of entities

Entity 1	Relationship	Entity 2	Description
Site	Inside	State	One or more sites are located inside each state
Site	Partitions into	Replication	A site is spatially partitioned into multiple replications
Entry	Tested at	Replication	An entry is tested in multiple replications
Environment	Affects	Replication	A set of environment variables affects the growing conditions in a replication. Currently, detailed environmental attributes (e.g., soil nutrients and temperature) are not recorded at the replication level and are therefore the same for each site. This database design aims to provide flexibility for future inclusion of fine-grained information.
Experiment	Includes	Entry	An experiment includes many entry numbers, one for each cultivar
Experiment	Records	Environment	A set of environment variables are recorded by an experiment
Sponsor	Supports	Entry	A sponsor funds one or more entries in an experiment
Cultivar	Represented by	Entry	A cultivar is tested as an entry in an experiment
Cultivar	Belongs to	Species	Each species has a set of cultivars that do not overlap with that of any other species
Cultivar	Commercially described by	Commercial_information	A cultivar is commercially described by commercial information
Entry	Tests	Rating	The ratings of different attributes are tested for each entry

subtypes are inherited from entities with identifying relationships and parents).

2.2.1 | Description of the ERD

An NTEP experiment is a 5-year process starting from the trial set year, and evaluates a set of turfgrass cultivars. Each cultivar is represented by a unique entry number in an experiment. To determine how well a cultivar is adapted across a number of environments, each is tested at multiple distinct sites within its geographic range of adaptation, with only a few states having multiple sites. A site is further partitioned into multiple replications and each cultivar is tested in all replications at each site, with all cultivars at a given site being tested under the same environmental

conditions.[1] Periodically (e.g., every month, season, year), researchers go through the sites to rate the quality, color and other performance attributes for all cultivars in all replications. The cost of testing a cultivar is covered by a sponsor (e.g., public plant breeding program, seed company), and the cultivar may be made commercially available for consumers depending on experimental results. Looking at some examples of cardinality constraints, we can see that a cultivar can only belong to one species while a species can have multiple cultivars. Similarly, a cultivar can only be supported by one sponsor in an experiment, but a sponsor can support multiple cultivars. Finally, note that the "plot" entity refers to an actual experimental unit (rectangular patch of turfgrass) inside a replication. The relative coordinates (row and column) of each plot for a given site may be useful in the future but is not yet recorded by NTEP, so we use a dashed green box to indicate this. This case also shows the benefit of an ERD in allowing a broader view of the data (e.g., future plans) rather than being limited by the existing data.

2.3 | Logical level design

The goal of logical level design is to map the Entity-Relationship diagram in a logical schema (i.e. database tables) using a set of conversion rules (Navathe & Elmasri, 2001), including those for mapping entities, relationships, etc. In the following, we describe conversion rules to map entities and relationships.

2.3.1 | Entity mapping

To map an entity from an ERD to a logical schema, we first create a table with all of the simple attributes of that entity. Then, we add a primary key constraint, which specifies the minimal set of attributes needed to uniquely identify each record (row) in the table. This constraint requires that no values of the attributes in the primary key be null, and that no two records have identical attribute values on the whole key. As an example, the entity "site" has four attributes: (state_id, site_id, site_name, and site_geometry), where the pair "state_id" and "site_id" is the primary key. Each table has one and only one primary key.

2.3.2 | Relationship mapping

Mapping relationships from ERD to logical database tables depends on relationship cardinality (e.g., one-to-one, one-to-many and many-to-many). Since the ERD in Figure 5 mainly involves a one-to-many (i.e., 1:N) relationship, we will use this as an example to illustrate the conversion rule – the **foreign key** approach. Denote T1 and T2 as two table schemas, where each record in T1 may correspond to many records in T2. A foreign key is always stored in T2, and it basically consists of the same set of attributes in T1's primary key; however, it may or may not be part of T2's primary key. The foreign key constraint requires that each combination of attribute values in the foreign key of T2 must exist in T1. In this way, each record in T2 can be connected to a specific record in T1 and many records in T2 may connect to the same records in T1. Figure 6 shows the mapping of two example relationships "inside" and "partitions into" excerpted from the ERD in Figure 5.

The complete set of logical database tables created using entity and relationship mapping is shown in Figure 7.

2.4 | Physical level design

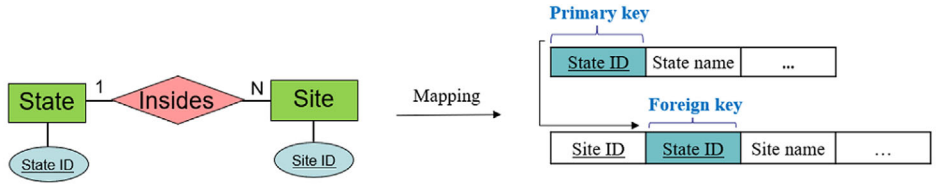
Physical level design concerns file organizations and related physical parameters for the database files. Its aim is to improve the storage and processing efficiency of the database. A key design decision at this level is the choice of database index, which is a precomputed search structure (e.g., through sorting) built for frequently used columns (e.g., primary key) to reduce the cost of query processing. In NTEP-DB-1.0, we built a B-tree index (Navathe & Elmasri, 2001; Shekhar & Chawla, 2003; Comer, 1979) for the primary key of each table as well as other frequently used attributes such as "quality_rating".

Finally, we implemented NTEP-DB-1.0 using PostgreSQL, one of the most popular open-source platforms for relational databases (Postgresql, 2020; Viloria et al., 2019). PostgreSQL is well-supported by its active community members, providing many resources to facilitate potential users (e.g., horticultural science researchers, seed companies, NTEP staff) in learning and using the database. Another advantage of Postgres is that it supports formal spatial data types (e.g., points, polygons, rasters defined by the Open-Geospatial-Commons' standards [OGC, 2020; PostGIS, 2020; OSGeo, 2020]), which NTEP could incorporate into its data collection in future experiment trials (detailed in the discussion).

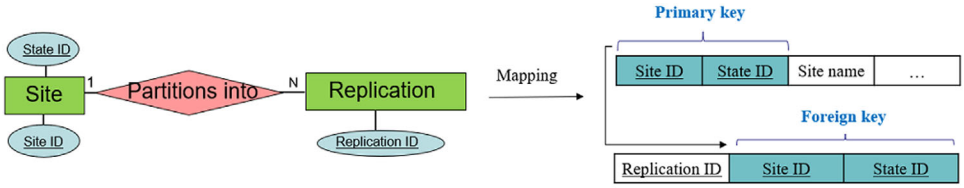
3 | RESULTS

The overall validation framework of NTEP-DB 1.0 is shown in Figure 8. Specifically, through the experiments, we aim to answer two main questions: (1) Sanity check: Do corresponding queries on NTEP-DB 1.0 return outputs that match existing summary reports on the NTEP website? (2) Flexibility

[1] Variations of environmental characteristics across replications are not recorded in existing field experiments. To enable future inclusion of more fine-grained information, the current database design allows replications to have different environmental attributes.



(a) Mapping of "Inside" relationship



(b) Mapping of "Partitions into" relationship

FIGURE 6 Mapping examples of 1:N relationships

check: Is the database able to answer user queries of different types (from surveys and user groups)?

In the following, we will first summarize the current data progress for NTEP-DB 1.0, and then show experiment results of sanity and flexibility check.

3.1 | NTEP data progress

We have acquired all the NTEP data for the fine fescues, a group of taxa similar in appearance possessing different stress tolerances;[2] this group of grasses is considered useful in landscapes where resource inputs (water, fertilizer, etc.) are limited (Braun et al., 2020). The first round of data insertion covered years from 1982 to 2016, and new field experiment data for 2017 was inserted in 2019.[3]

Fine fescue data were provided by NTEP to test this first version of the database. In 2019, the database design and query results were presented to the NTEP board during its annual advisory board meeting, and the board has now voted and agreed to move forward and put data for all tested turfgrass species into NTEP-DB 1.0.

3.2 | Sanity check: A comparison with NTEP summary reports

First, we performed a sanity check to confirm that the results from the NTEP database were correct. We used a set of existing summary reports from the NTEP website (NTEP, 2020a) as the ground truth to see if the outputs from the database queries can match the report data.

Figure 9 shows the comparison for green-up ratings for Chewings fescue cultivars in 2012. Figure 9a shows the screenshot of the summary report from the NTEP website (NTEP, 2012) and Figure 9b shows the corresponding query output from the NTEP database. The results of the query were put into the same order as the summary report to make the comparison easier. The first column in the table lists the names of the cultivars, and the rest of the columns show the statistics (i.e., mean green-up rating) recorded at different test sites in different states (the name of a test site is listed as the abbreviation of the state followed by the ID of the test site within that state). As we can see, for this comparison, the statistics from the query output (Figure 9b) match those in the summary report (Figure 9a). The same trend can be seen in Figures 10 and 11, which show the same type of comparison for Chewings fescue cultivars in 2009 and 2007. These results help validate that the data were correctly inserted into the database.[4]

[2] Taxa: Strong creeping red fescue (*F. rubra* L. ssp. *rubra* Gaudin), slender creeping red fescue [*F. rubra* L. ssp. *littoralis* (G. Mey.) Auquier], Chewings fescue [*F. rubra* L. ssp. *commutata* Gaudin; syn. *F. rubra* L. ssp. *fallax* (Thuill.) Nyman], hard fescue (*F. brevipila* Tracey), and sheep fescue [*F. ovina* L.; syn. *F. ovina* L. ssp. *hirtula* (Hack. ex Travis) M.J. Wilk].

[3] It currently takes some time for NTEP to gather and process raw data from different participating institutions before making them ready for insertion into the database.

[4] Small differences in means found on the NTEP website versus the database query are either due to (1) rounding, where a 0.1 difference is noted from the web site and database query values, or (2) adjusted means on the web site data resulting from the use of a different statistical procedure than means computed by the database query (https://ntep.org/LPI%20reporting%20Q&A%205-9-13.pdf).

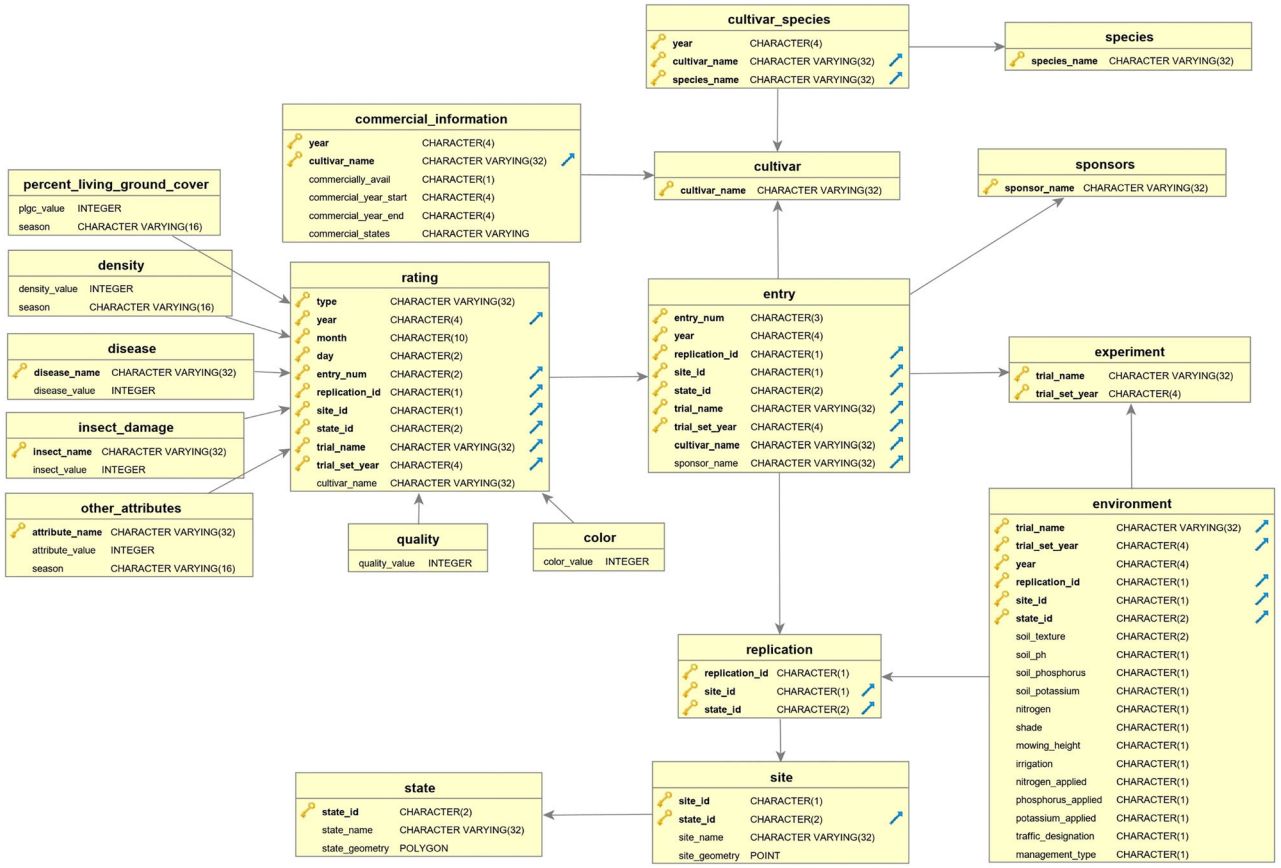


FIGURE 7 NTEP database with complete list of attributes corresponding to each table

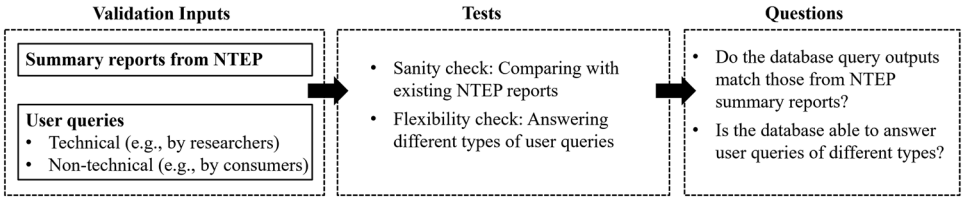


FIGURE 8 Overall validation framework and questions

NAME	IL1	NE1	RI1	UT1	WA3	MEAN
PST-4CSD	4.3	5.0	5.7	4.0	6.0	5.0
CASCADE	3.3	5.0	5.3	4.3	6.0	4.8
INTRIGUE 2	5.3	4.0	4.7	4.3	5.0	4.7
FAIRMONT (TCD)	4.3	4.7	3.7	4.3	6.0	4.6
WRIGLEY 2 (IS-FRC-35)	4.0	4.7	4.3	4.7	5.3	4.6
ZODIAC	4.3	4.3	3.7	4.7	5.7	4.5
LACROSSE	3.7	4.7	4.3	4.0	5.3	4.4
RADAR (MVS-FRC-101)	3.7	4.7	3.7	3.7	6.3	4.4
TREAZURE II	5.0	4.3	4.3	2.7	5.7	4.4
LONGFELLOW 3 (IS-FRC-33)	4.0	4.0	4.0	4.3	5.3	4.3
PSG 50C3	3.7	4.3	3.7	4.0	6.0	4.3

(a) Summary report from NTEP website

Cultivar_name	IL1	NE1	RI1	UT1	WA3	MEAN
Shadow III (PST-4CSD)	4.33	5	5.67	4	6	5.00
Cascade	3.33	5	5.33	4.33	6	4.80
Intrigue 2	5.33	4	4.67	4.33	5	4.67
Fairmont (TCD)	4.33	4.67	3.67	4.33	6	4.60
Wrigley 2 (IS-FRC-35)	4	4.67	4.33	4.67	5.33	4.60
Zodiac	4.33	4.33	3.67	4.67	5.67	4.53
Lacrosse	3.67	4.67	4.33	4	5.33	4.40
Radar (MVS-FRC-101)	3.67	4.67	3.67	3.67	6.33	4.40
Treazure II	5	4.33	4.33	2.67	5.67	4.40
Longfellow 3 (IS-FRC-33)	4	4	4	4.33	5.33	4.33
PSG 50C3	3.67	4.33	3.67	4	6	4.33

(b) Output of NTEP database query

FIGURE 9 Test-1: Spring green-up ratings of Chewings fescue cultivars in 2012 (NTEP, 2012)

NAME	RI1	VA1	WI1	MEAN
IS-FRC-33	8.1	7.5	7.3	7.6
MVS-FRC-101	8.1	7.4	7.1	7.5
FAIRMONT (TCD)	7.8	7.4	6.8	7.3
ZODIAC	7.2	7.3	7.1	7.2
INTRIGUE 2	7.2	7.4	6.9	7.2
TREAZURE II	7.3	7.1	6.6	7.0
IS-FRC-35	7.0	7.2	6.8	7.0
PSG 50C3	6.9	7.1	6.2	6.7
LACROSSE	6.7	6.8	5.6	6.4
CASCADE	5.7	6.7	5.7	6.0
PST-4CSD	4.8	6.1	4.6	5.1

(a) Summary report from NTEP website

Cultivar_name	RI1	VA1	WI1	Mean
Longfellow 3 (IS-FRC-33)	8	7.48	7.53	7.67
Radar (MVS-FRC-101)	8.2	7.62	7.07	7.63
Fairmont (TCD)	8	7.35	6.33	7.23
Zodiac	7.6	6.5	7.13	7.08
Intrigue 2	7.8	7.5	6.53	7.28
Wrigley 2 (IS-FRC-35)	7.2	7.1	6.53	6.94
Treazure II	7	6.76	6.93	6.90
PSG 50C3	6.4	7.76	6.07	6.74
Lacrosse	6.6	7.14	5.87	6.54
Cascade	6	6.9	5.2	6.03
Shadow III (PST-4CSD)	5.6	6.43	4.2	5.41

(b) Output of NTEP database query

FIGURE 10 Test-2: Quality ratings of Chewings fescue cultivars in 2009 (NTEP, 2009)

NAME	MA1	NJ1	NJ2	NY1	PA1	QE1	RI1	MEAN
ZODIAC (BUR 4601)	5.1	6.6	5.4	4.7	7.1	5.9	7.0	6.0
SR 5130 (SRX 51G)	4.2	4.6	5.8	4.9	6.8	6.1	7.5	5.7
TREAZURE II (PST-4TZ)	5.5	3.9	5.1	5.7	5.9	6.1	7.2	5.6
DP 77-9885	4.4	5.5	5.1	4.9	6.8	5.7	6.8	5.6
COMPASS (ACF 188)	3.8	5.2	5.3	5.3	6.3	5.7	6.9	5.5
LACROSSE (IS-FRC 17)	3.5	5.3	5.5	5.1	6.3	5.6	7.1	5.5
LONGFELLOW II	3.6	5.1	5.6	5.0	6.1	5.9	7.0	5.5
7 SEAS	3.8	5.3	5.0	5.1	6.3	5.7	6.9	5.4
AMBASSADOR	3.9	4.6	5.3	5.0	5.8	5.9	7.1	5.4
DP 77-9886	3.9	4.5	4.5	5.1	5.6	5.7	7.0	5.2
MUSICA	4.1	2.7	4.6	5.0	5.9	5.6	7.1	5.0
J-5 (JAMESTOWN 5)	3.7	4.1	4.6	5.3	5.1	6.0	6.1	5.0
CULUMBRA II (ACF 174)	3.2	3.2	4.4	5.3	5.5	5.7	5.8	4.7
CASCADE	3.2	3.3	3.8	5.0	5.0	5.6	5.9	4.6

(a) Summary report from NTEP website

cultivar_name	MA1	NJ1	NJ2	NY1	PA1	QE1	RI1	Mean
Zodiac (BUR 4601)	5.48	6.62	5.42	4.71	7.11	7.11	7.05	6.21
SR 5130 (SRX 51G)	5.14	4.57	5.83	4.90	6.78	6.78	7.48	5.93
DP 77-9885	4.43	5.52	5.13	4.90	6.83	6.83	6.76	5.77
Treasure II (PST-4TZ)	4.19	3.86	5.13	5.71	5.89	5.89	7.19	5.41
Compass (ACF 188)	4.14	5.19	5.33	5.33	6.28	6.28	6.90	5.64
LaCrosse (IS-FRC 17)	3.90	5.29	5.50	5.10	6.28	6.28	7.10	5.64
Longfellow II	3.90	5.10	5.58	4.95	6.11	6.11	7.05	5.54
7 Seas	3.81	5.33	4.96	5.14	6.28	6.28	6.86	5.52
Ambassador	3.76	4.62	5.33	4.95	5.83	5.83	7.10	5.35
DP 77-9886	3.67	4.52	4.50	5.10	5.61	5.61	6.95	5.14
Musica	3.57	2.67	4.63	5.00	5.94	5.94	7.10	4.98
J-5 (Jamestown 5)	3.48	4.10	4.58	5.33	5.11	5.11	6.14	4.84
Culumbra II (ACF 174)	3.24	3.19	4.38	5.33	5.50	5.50	5.81	4.71
Cascade	3.24	3.33	3.75	5.05	5.00	5.00	5.90	4.47

(b) Output of NTEP database query

FIGURE 11 Test-3: Quality ratings of Chewings fescue cultivars in 2007 (NTEP, 2007)

3.3 | Flexibility: Query formulations and results based on user requests

We also worked with horticultural researchers to design a set of concrete queries based on the need of turfgrass professionals as well as the desired types of queries identified from earlier survey results (Yue et al., 2019). The set covers a variety of query types such as filtering, aggregation as well as ranking, and can help validate that the database design offers the flexibility needed to answer different types of queries.

In the following, we start from the query example used in the introduction (i.e., the one requiring hundreds of manual clicks in the NTEP website), and then move into five more complicated query examples. Note that in the NTEP data many attribute values are coded as integers, so each example has a brief explanation of the codes used in the query. A full code list is available from the NTEP website (NTEP, 2020b).

Query 1: List the names of fine fescue cultivars with an average quality rating of at least 6 from 2004 to 2014 in Minnesota.

Explanation: Quality rating is the visual estimate integrating all factors of turfgrass quality, and it is on a scale of 1 to 9 where 1 means grass is dead and 9 is maximum healthy turf. Note that currently Zodiac is listed twice in the result ("Zodiac" and "Zodiac (BUR 4601)"), which is potentially caused by inconsistent naming patterns used in different field experiments. Future versions may include additional preprocessing steps or constraints (e.g., together with the data collector in Sec. 4.3) to address this issue.

SQL code snippet and result: Figure 12a and b.

Query 2: Which fine fescue cultivar performs well in soil with pH around 6.5, shaded conditions with turf quality above 5?

Explanation: The codes for pH and shade are as follows (format: code - meaning):

- Soil pH (sample from 0–3 inches of depth): 1, 3.5 or less; 2, 3.6–4.5; 3, 4.6–5.5; 4, 5.6–6.0; 5, 6.1–6.5; 6, 6.6–7.0; 7, 7.1–7.5; 8, 7.6–8.5; and 9, 8.6 or greater.
- Shade: 1, Dense shade; 3, Uniform or artificial shade; 5, Partial shade; 7, Light shade; and 9, Full sun.

SQL code snippet and result: Figure 13a and b.

```
SELECT e.cultivar_name, AVG(q.quality_value) AS quality_avg
FROM entry e, environment env, quality q, State s
WHERE e.replication_id = env.replication_id AND e.year = env.year AND e.state_id = env.state_id
    AND e.site_id = env.site_id AND e.trial_set_year = env.trial_set_year AND e.trial_name = env.trial_name
    AND e.entry_num = q.entry_num AND e.replication_id = q.replication_id  AND e.year = q.year
    AND e.state_id = q.state_id AND e.site_id = q.site_id AND e.trial_set_year = q.trial_set_year
    AND e.trial_name = q.trial_name
    AND e.year >= '2004' AND e.year<= '2014' AND s.state_id = e.state_id  AND state_name = 'Minnesota'
GROUP BY e.cultivar_name
HAVING AVG(Q.quality_value)>=6
ORDER BY AVG(quality_value) DESC;
```

(a)

	cultivar_name character varying (32)	quality_avg numeric
1	Fairmont (TCD)	6.6363636363636364
2	PSG 50C3	6.5075757575757576
3	Radar (MVS-FRC-101)	6.3181818181818182
4	Treazure II	6.3030303030303030
5	Spartan II (Pick HF #2)	6.2173913043478261
6	Longfellow 3 (IS-FRC-33)	6.1893939393939394
7	Zodiac	6.1363636363636364
8	Intrigue 2	6.0833333333333333
9	PST-8000	6.0289855072463768
10	Zodiac (BUR 4601)	6.0147058823529412

(b)

FIGURE 12 (a) SQL code for Query 1 – "List the names of fine fescue cultivars with an average quality rating of at least 6 from 2004 to 2014 in Minnesota"; (b) Results in a descending order

```
SELECT e.cultivar_name, AVG(q.quality_value) AS quality_avg
FROM entry e, quality q, environment env
WHERE e.replication_id = env.replication_id AND e.year = env.year AND e.state_id = env.state_id
     AND e.site_id = env.site_id AND e.trial_set_year = env.trial_set_year AND e.trial_name = env.trial_name
     AND e.entry_num = q.entry_num AND e.replication_id = q.replication_id AND e.year = q.year
     AND e.state_id = q.state_id AND e.site_id = q.site_id AND e.trial_set_year = q.trial_set_year
     AND e.trial_name = q.trial_name
     AND soil_ph <= '7'  AND soil_ph <> ' ' AND soil_ph >= '6' AND shade <= '5'
GROUP BY e.cultivar_name
HAVING AVG(q.quality_value)>5
ORDER BY AVG(q.quality_value) DESC;
```

(a)

	cultivar_name character varying (32)	quality_avg numeric
1	SR 5130 (SRX 51G)	7.2028985507246377
2	Waldina	6.9166666666666667
3	Spartan II (Pick HF #2)	6.8695652173913043
4	PPG-FRT 101	6.8472222222222222
5	4LS (Bighorn)	6.7500000000000000
6	Reliant IV (A01630Rel)	6.6521739130434783
7	Berkshire	6.5942028985507246
8	DLF-RCM	6.5942028985507246
9	BAR Fo 81-225	6.5833333333333333
10	PST-4BEN	6.5833333333333333

(b)

FIGURE 13 (a) SQL code for Query 2 – "Which fine fescue cultivar performs well in soil with pH around 6.5, shaded conditions with turf quality above 5?"; and (b) Top 10 results of best-performing cultivars out of 196 returned results

Query 3: Which commercially available fine fescue cultivars perform best in October in Minnesota?

Explanation: Only a subset of cultivars tested in field trials are commercially available.

SQL code snippet and result: Figure 14a and b.

Query 4: Which entries by location and year performed better than or equal to 6.0 when maintained with nitrogen inputs of less than 2.0 lb of nitrogen per 1000 square feet per year?

Explanation: The codes for nitrogen levels are as follows (unit is pounds of nitrogen per 1000 square feet per year; format: code - range): 1, 0–1.0; 2, 1.1–2.0; 3, 2.1–3.0; 4, 3.1–4.0; 5, 4.1–5.0; 6, 5.1–6.0; 7, 6.1–7.0; 8, 7.1–8.0; and 9, 8.1 or greater.

SQL code snippet and result: Figure 15a and b.

Query 5: Which Chewings fescue cultivars had the worst percent ground cover with mechanical traffic?

Explanation: Codes for traffic designations by mechanical type: 1, No traffic; 2, Spring; 3, Summer; 4, Fall; 5, Winter; and by athletic type: 6, Spring; 7, Summer; 8, Fall; 9, Winter.

SQL code snippet and result: Figure 16a and b.

Query 6: Which fine fescue cultivars had average turf quality greater than 6.0 for July, August and September when mowed at a height over 2.0 inches?

Explanation: Codes for mowing height are as follows: 1, 0–0.5"; 2, 0.6–1.0"; 3, 1.1–1.5"; 4, 1.6–2.0"; 5, 2.1–2.5"; 6, 2.6–3.0"; 7, 3.1–3.5"; 8, 3.6–4.0"; and 9, 4.1 or greater.

SQL code snippet and result: Figure 17a and b.

4 | DISCUSSION

In this section, we discuss three recommendations we have for the next generation of the NTEP database: a web interface, real-time data update, and a spatial database.

4.1 | A web interface for non-technical users and beyond

Since consumers (e.g., homeowners, sales people) are unlikely to have database knowledge needed to perform

```sql
SELECT distinct e.cultivar_name, AVG(q.quality_value) AS quality_avg, COUNT(q.quality_value) AS count_readings
FROM entry e, quality q, environment env, state s, commercial_information c
WHERE e.replication_id = env.replication_id AND e.year = env.year AND e.state_id = env.state_id
    AND e.site_id = env.site_id AND e.trial_set_year = env.trial_set_year AND e.trial_name = env.trial_name
    AND e.entry_num = q.entry_num AND e.replication_id = q.replication_id AND e.year = q.year
    AND e.state_id = q.state_id AND e.site_id = q.site_id AND e.trial_set_year = q.trial_set_year
    AND e.trial_name = q.trial_name
    AND e.state_id = s.state_id
    AND e.cultivar_name = c.cultivar_name AND e.year = c.year
    AND c.commercially_avail = 'Y' AND q.month in ('october')
    AND s.state_name = 'Minnesota'
GROUP BY e.cultivar_name
ORDER BY AVG(q.quality_value) DESC;
```

(a)

	cultivar_name character varying (32)	quality_avg numeric
1	Fairmont (TCD)	7.2000000000000000
2	Shoreline (SRX 55R)	6.6666666666666667
3	Gladiator (TH456)	6.5333333333333333
4	Berkshire	6.3333333333333333
5	Spartan II (Pick HF #2)	6.3333333333333333
6	Beacon	6.2666666666666667
7	Radar (MVS-FRC-101)	6.2666666666666667
8	Sword	6.2666666666666667
9	Longfellow 3 (IS-FRC-33)	6.2000000000000000
10	Radar	6.2000000000000000

(b)

FIGURE 14 (a) SQL code for Query 3 – "Which commercially-available fine fescue cultivars perform best in October in Minnesota?"; and (b) Top 10 results of best-performing cultivars out of 68 returned results

```sql
SELECT DISTINCT e.cultivar_name, e.year, s.state_name,
AVG(q.quality_value) AS quality_avg, COUNT(q.quality_value) AS count_readings
FROM entry e, quality q, environment env, state s
WHERE e.replication_id = env.replication_id AND e.year = env.year AND e.state_id = env.state_id
    AND e.site_id = env.site_id AND e.trial_set_year = env.trial_set_year AND e.trial_name = env.trial_name
    AND e.entry_num = q.entry_num AND e.replication_id = q.replication_id AND e.year = q.year
    AND e.state_id = q.state_id AND e.site_id = q.site_id AND e.trial_set_year = q.trial_set_year
    AND e.trial_name = q.trial_name AND e.state_id = s.state_id
    AND env.nitrogen<= '2' AND env.nitrogen <> ' '
GROUP BY e.cultivar_name, e.year, s.state_name
HAVING MIN(q.quality_value)>= 6
ORDER BY AVG(q.quality_value) DESC;
```

(a)

	cultivar_name character varying (32)	year character (4)	state_name character varying (32)	quality_avg numeric
1	Gotham (IS-FL 28)	2006	Iowa	8.6111111111
2	PSG 50C3	2012	Rhode Island	8.5833333333
3	Predator	2006	Iowa	8.5555555555
4	Spartan II (Pick HF #2)	2006	Iowa	8.5555555555
5	Radar (MVS-FRC-101)	2012	Rhode Island	8.5416666666
6	Berkshire	2006	Iowa	8.5000000000
7	Epic (5001)	2006	Wisconsin	8.5000000000
8	Reliant IV (A01630Rel)	2006	Iowa	8.4444444444
9	Gotham (IS-FL 28)	2005	Wisconsin	8.3888888888
10	Radar (MVS-FRC-101)	2011	Rhode Island	8.3750000000

(b)

FIGURE 15 (a) SQL code for Query 4 – "Which entries by location and year performed better than or equal to 6.0 when maintained with nitrogen inputs of less than 2.0 lb of nitrogen per 1000 square feet per year?"; and (b) Top 10 results of best-performing cultivars out of 603 returned results

queries, we will design and implement an easy-to-use webapp for non-technical users that can be used on a variety of devices (e.g., smartphones, tablets, desktops). The application can be considered as an interface between users and the NTEP database, allowing users such as homeowners to get recommendations of turfgrass cultivars to purchase by just opening the app while standing on their lawn or answering a minimal set of simple questions that do not require any expert knowledge.

Figure 12 shows a prototype of the web application's interface for homeowners on a smartphone. As shown in the figure, homeowners can use the app to describe their lawn and

```sql
SELECT e.cultivar_name, AVG(q.plgc_value)
FROM entry e, percent_living_ground_cover q, environment env, cultivar_species c
WHERE e.replication_id = env.replication_id AND e.year = env.year AND e.state_id = env.state_id
    AND e.site_id = env.site_id AND e.trial_set_year = env.trial_set_year AND e.trial_name = env.trial_name
    AND e.entry_num = q.entry_num AND e.replication_id = q.replication_id AND e.year = q.year
    AND e.state_id = q.state_id AND e.site_id = q.site_id AND e.trial_set_year = q.trial_set_year
    AND e.trial_name = q.trial_name AND e.cultivar_name = c.cultivar_name AND e.year = c.year
    AND env.traffic_designation = '3' AND c.species_name IN ('chewing', 'chewings')
GROUP BY e.cultivar_name
ORDER BY AVG(q.plgc_value);
```

(a)

	cultivar_name character varying (32)	avg numeric
1	Culumbra II (ACF 174)	87.7777777777777778
2	Musica	90.5555555555555556
3	Cascade	91.4444444444444444
4	7 Seas	92.4444444444444444
5	SR 5130 (SRX 51G)	92.8888888888888889
6	Compass (ACF 188)	93.1111111111111111
7	Longfellow II	94.0000000000000000
8	DP 77-9886	94.2222222222222222
9	Treasure II (PST-4TZ)	94.2222222222222222
10	DP 77-9885	94.5555555555555556

(b)

FIGURE 16 (a) SQL code for Query 5 – "Which Chewings fescue cultivars had the worst percent ground cover with mechanical traffic?"; and (b) Top 10 results of best-performing cultivars out of 14 returned results

```
SELECT DISTINCT e.cultivar_name, avg(q.quality_value) AS quality_avg
FROM entry e, quality q, environment env
WHERE e.replication_id = env.replication_id AND e.year = env.year AND e.state_id = env.state_id
    AND e.site_id = env.site_id AND e.trial_set_year = env.trial_set_year AND e.trial_name = env.trial_name
    AND e.entry_num = q.entry_num AND e.replication_id = q.replication_id AND e.year = q.year
    AND e.state_id = q.state_id AND e.site_id = q.site_id AND e.trial_set_year = q.trial_set_year
    AND e.trial_name = q.trial_name
    AND env.mowing_height>='4' AND q.month IN ('july', 'august', 'september')
GROUP BY e.cultivar_name
HAVING AVG(q.quality_value)>=6
ORDER BY AVG(q.quality_value) DESC;
```

(a)

	cultivar_name character varying (32)	quality_avg numeric
1	Spartan II (Pick HF #2)	6.2204585537918871
2	Warwick	6.1317829457364341
3	Reliant IV (A01630Rel)	6.1111111111111111
4	Epic (5001)	6.1093474426807760
5	Gotham (IS-FL 28)	6.0987654320987654
6	Zodiac (BUR 4601)	6.0229681978798587

(b)

FIGURE 17 (a) SQL code for Query 6 – "Which Chewings fescue cultivars had the worst percent ground cover with mechanical traffic during summer?"; and (b) Results sorted in a descending order by rating

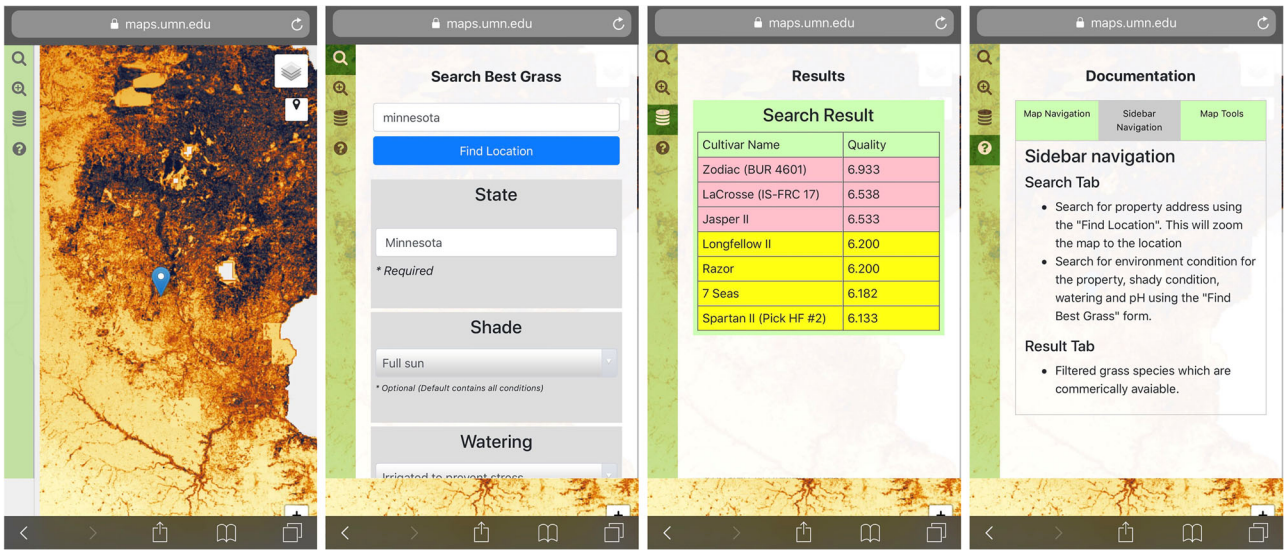


FIGURE 18 A prototype of the webapp on a mobile device

management conditions such as amount of shading, soil type, irrigation needs, mowing regiment and use of the lawn. These inputs are then mapped to variables in the NTEP database, and using a pre-formatted SQL query described above, grass seed with the highest quality score is recommended. As NTEP trial locations are distributed across the country, this query approach not only looks at the trial site closest to the homeowner, but also sites that have similar climate or are located in the same ecoregion. A map interface can also be used to locate the area of interest, which can minimize user inputs by automatically filling in information such as shade conditions, soil types and additional climate characteristics. Furthermore, beyond helping consumers such as homeowners, we also aim to design and develop functionalities or webapps to support other common use cases of the NTEP database, including:

- Canned reports for industry professionals and researchers;
- Guided queries for professionals and researchers;
- An API for advanced users wanting to define their own queries.

4.2 | Real-time data updating with an in-field data collector

One major issue with the NTEP database is that data collection takes many steps over a long period: (1) Data collection in the field is done by individual institutions and labs in their own formats (e.g., hand-written, collected on a mobile device); (2) Field recordings are manually converted into digital formats, post-processed and submitted to NTEP; (3) NTEP gathers the datasets from the various sources and merges them into a single format; (4) The integrated data are uploaded to the database.

This process is not only long and tedious, but also greatly increases the risk of data quality issues such as inconsistency and errors. For example, some common issues

we observed during the data insertion process included data value violations (e.g., out-of-range data values, or letters written in numeric fields), inconsistencies in naming (e.g., species "slender creeping red" written as "slender creep", "slender creeping" and others), missing values, etc. The more the errors accumulate, the more challenging and time-consuming it becomes to identify and fix them.

To address this issue once and for all, we recommend completely overhauling the data collection process and developing new data collector applications that can directly gather experiment information in the field in digital format and automatically upload it to the database. This would also allow the database to automatically check the validity of the input against a predefined set of rules, and avoid or reduce the data quality issues.

4.3 | Towards a spatial database

The current NTEP database does not contain explicit spatial information (e.g., geo-coordinates, coordinate system, projection), and only has state names and site IDs within states that can be used to infer high-level geographic regions. While spatial entities were included in the conceptual design of the database (Figure 5), they were mainly incorporated for ease of extension in the future and such information has not yet been collected in NTEP experiments. For example, neither the spatial coordinates of replications and plots nor their mutual spatial relationships have been recorded. However, fine-scale spatial information can be useful for analyzing field trials since variability in soils and other environmental characteristics can be problematic.

Furthermore, spatial information also allows the potential use of advanced spatial data science techniques (Atluri et al., 2018; Xie et al., 2017) such as spatial pattern mining (e.g., detecting hotspots of disease or stress; finding co-locations between a disease and other environmental conditions), learning (e.g., quality rating prediction with deep learning [Skakun et al., 2017]) and optimization (e.g., improved allocation and management design for field experiments).

Fortunately, spatial databases and design techniques such as spatial pictograms (Shekhar & Chawla, 2003) have matured to provide functionalities in storing and managing spatial information (e.g., plot polygons, hydrological network), and there is a tremendous amount of spatial data (e.g., high-resolution topographical models and slope information derived from LiDAR point clouds) that can be leveraged to enrich the NTEP database to support advanced data analytics.

5 | CONCLUSIONS AND FUTURE WORK

We carried out a multi-disciplinary and multi-sectoral work to create NTEP-DB 1.0, the first version of the database for the National Turfgrass Evaluation Program. Specifically, we presented the need for the database and proposed conceptual (e.g., ER diagram) and logical level designs. We validated the design through an implementation of the database in Postgres. The experiments showed that the outputs are correct and the database is flexible in answering various types of user queries. NTEP-DB 1.0 is a milestone achievement for both NTEP and the turfgrass communities that use it.

To further advance the field, we also provided three recommendations for the next generation of the database: a user-database interface/webapp, real-time in-the-field data collector, and a spatial database. We plan to continue to investigate these two recommendations and incorporate them into the next version. Additionally, in the short term, we will work with horticultural science researchers as well as NTEP managers to develop a set of rules needed to address the data quality issues (e.g., naming inconsistency). In the long term, we will explore new spatial data science techniques that can potentially advance NTEP data analytics.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants No. 1737633, 1901099, 1916518, and 2105133, the USDA under Grant No. 2017-51181-27222, and the OVPR Infrastructure Investment Initiative at the University of Minnesota.

ORCID

Len Kne https://orcid.org/0000-0003-1932-3101

REFERENCES

Atluri, G., Karpatne, A., & Kumar, V. (2018). Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys (CSUR)*, *51*, 1–41. https://doi.org/10.1145/3161602

Braun, R. C., Patton, A. J., Watkins, E., Koch, P. L., Anderson, N. P., Bonos, S. A., & Brilman, L. A. (2020). Fine fescues: A review of the species, their improvement, production, establishment, and management. *Crop Science*, *60*, 1142–1187. https://doi.org/10.1002/csc2.20122

Comer, D. (1979). Ubiquitous B-tree. *ACM Computing Surveys (CSUR)*, *11*, 121–137. https://doi.org/10.1145/356770.356776

Garcia-Molina, H., Ullman, J. D., & Widom, J. (2000). *Database system implementation*. Upper Saddle River, NJ: Prentice Hall.

Li, Q., & Chen, Y. L. (2009). Entity-relationship diagram. In *Modeling and analysis of enterprise and information systems* (pp. 125–139). Berlin, Heidelberg: Springer.

Mannino, M. V. (2005). *Database design, application development, and administration*. McGraw-Hill, Inc.

Morris, K. N., & Shearman, R. C. (2000). The National Turfgrass Evaluation Program: Assessing new and improved turfgrasses. *Diversity*, *16*, 19–22.

National Turfgrass Evaluation Program (NTEP). (2007). Mean turfgrass quality ratings of Chewings fescue cultivars. http://www.ntep.org/data/ff03/ff03_08-6/ff0308t01b.txt

National Turfgrass Evaluation Program (NTEP). (2009). Turfgrass quality ratings of Chewings fescue cultivars. http://www.ntep.org/data/ff08/ff08_10-3/ff0810t02b.txt

National Turfgrass Evaluation Program (NTEP). (2012). Spring greenup ratings of Chewings fescue cultivars. http://www.ntep.org/data/ff08/ff08_13-5/ff0813t09b.txt

National Turfgrass Evaluation Program (NTEP). (2020a). https://www.ntep.org/

National Turfgrass Evaluation Program (NTEP). (2020b). Code List. https://www.ntep.org/pdf/codelist.pdf

Navathe, S. B., & Elmasri, R. A. (2001). *Fundamentals of database systems with cdrom and book*. Addison-Wesley Longman Publishing Co., Inc.

OGC. (2020). OGC: Open Geospatial Commons. https://www.ogc.org/

OSGeo on PostGIS. (2020). PostGIS Spatial Database Extension. https://www.osgeo.org/projects/postgis/

PostGIS. (2020). PostGIS: Spatial and Geographic objects for PostgreSQL. https://postgis.net/

Postgresql. (2020). Postgresql: The World's Most Advanced Open Source Database. https://www.postgresql.org/

Skakun, S., Vermote, E., Roger, J. C., & Franch, B. (2017). Combined use of Landsat-8 and Sentinel-2A images for winter crop mapping and winter wheat yield assessment at regional scale. *AIMS geosciences*, *3*, 163. https://doi.org/10.3934/geosci.2017.2.163

Shekhar, S., & Chawla, S. (2003). *Spatial Databases: A Tour*. Pearson.

Viloria, A., Acuña, G. C., Alcázar Franco, D. J., Hernández-Palma, H., Fuentes, J. P., & Rambal, E. P. (2019). Integration of data mining techniques to PostgreSQL database manager system. *Procedia Computer Science*, *155*, 575–580. https://doi.org/10.1016/j.procs.2019.08.080

Xie, Y., Eftelioglu, E., Ali, R., Tang, X., Li, Y., Doshi, R., & Shekhar, S. (2017). Transdisciplinary foundations of geospatial data science. *ISPRS International Journal of Geo-Information*, *6*, 395. https://doi.org/10.3390/ijgi6120395

Yue, C., Wang, J., Watkins, E., Xie, Y., Shekhar, S., Bonos, S. A., Patton, A., Morris, K., & Moncada, K. (2019). User preferences for accessing publically available turfgrass cultivar performance data. *HortTechnology*, *29*, 599–610. https://doi.org/10.21273/HORTTECH04390-19

Yue, C., Wang, J., Watkins, E., Bonos, S. A., Nelson, K. C., Murphy, J. A., Meyer, W. A., & Horgan, B. P. (2017). Heterogeneous consumer preferences for turfgrass attributes in the United States and Canada. *Canadian Journal of Agricultural Economics*. https://doi.org/10.1111/cjag.12128

How to cite this article: Xie Y, Farhadloo M, Guo N, et al. NTEP-DB 1.0: A relational database for the national turfgrass evaluation program. *Int Turfgrass Soc Res J*. 2022;*14:*316–332. https://doi.org/10.1002/its2.76