CONSOLIDACIÓN DE LOS DATOS HISTÓRICOS DEL PROGRAMA DE MEJORAMIENTO GENÉTICO DE ARROZ EN UNA BASE DE DATOS

I. Rebollo[1], S. Scheffel[1], W. Iriarte[2], P. Blanco[3], F. Molina[4], F. Pérez de Vida[5], J. E. Rosas[6]

PALABRAS CLAVE: análisis conjunto, gestión de la información, integración de datos.

INTRODUCCIÓN

El Programa de Mejoramiento Genético de Arroz de INIA (PMGA) genera cada año más de 95.000 datos correspondientes a casi 7000 registros de ensayos de evaluación de rendimiento y análisis de laboratorio. Estos se archivan separadamente por ensayo, sin integrarse a una base de datos común, lo que dificulta su acceso y análisis en conjunto. El volumen de datos generado, así como la necesidad de capitalizar los avances en modelado estadístico y herramientas genómicas e integrarlos en su rutina de funcionamiento, hacen que el PMGA requiera adecuar su sistema de manejo de la información (Odell *et al*. 2017). El objetivo de este trabajo, en el marco del Proyecto ANII FSDA_1_2018_1_154120, fue la integración de toda la información disponible generada por el PMGA en una base de datos relacional, para su uso y actualización eficiente.

MATERIALES Y MÉTODOS

Se recopilaron los registros disponibles desde 1997 hasta la fecha de los ensayos de campo, cama de infección de enfermedades, y laboratorios de calidad industrial y culinaria, para la evaluación de líneas fijas de los subprogramas que integran el PMGA (*índica*, *japónica* tropical, *japónica* templado, Clearfield, e Híbridos), en las etapas de evaluación 1 a 5 y Final. Las planillas de los subprogramas *japónica* tropical, Clearfield e Híbridos eran de dos tipos: tipo A con la ID de las líneas experimentales y su pedigrí, y tipo B con las variables fenotípicas medidas en el ensayo, para ellas se aplicó la metodología esquematizada en figura 1.1. Brevemente, se crearon y aplicaron códigos en R para leer y fusionar ambos tipos de planilla en un único *data frame* por ensayo, colapsándose todas las columnas que correspondieran a la misma variable. Para los datos de los subprogramas *índica* y *japónica* templado (Figura 1.2) se diseñó una planilla de formato uniforme con nombres estandarizados para los campos, creándose una planilla por ensayo. Éstas se recopilaron mediante un código en R en un único *data frame*. Posteriormente, se unificaron todos los *data frames* en uno con toda la información del PMGA, y se estandarizaron valores para datos perdidos y niveles de variables categóricas. Se aplicó un control de calidad con criterios estadísticos y agronómicos. Se separaron los datos en planillas correspondientes a las tablas para la creación de la base de datos, definiéndose la estructura de la base de datos relacional según los elementos que constituyen el sistema y sus relaciones. Las planillas se migraron a una base de datos gestionada mediante sistema SQL, formada por tablas estructuradas en registros y campos vinculados por un identificador.

[1] Ing. Agr. INIA. Estudiante de maestría Cs. Agrarias F. Agronomía, UdelaR.
[2] Br. INIA Unidad de Biotecnología
[3] Ing. Agr. M.Sc. INIA. Programa Nacional de Investigación en Producción de Arroz, hasta junio 2018
[4] Ing. Agr. Ph.D. INIA. Programa Nacional de Investigación en Producción de Arroz
[5] Ing. Agr. M.Sc. Ph.D. INIA. Programa Nacional de Investigación en Producción de Arroz
[6] Lic. M.Sc. Dr. INIA. Programa Nacional de Investigación en Producción de Arroz jrosas@inia.org.uy

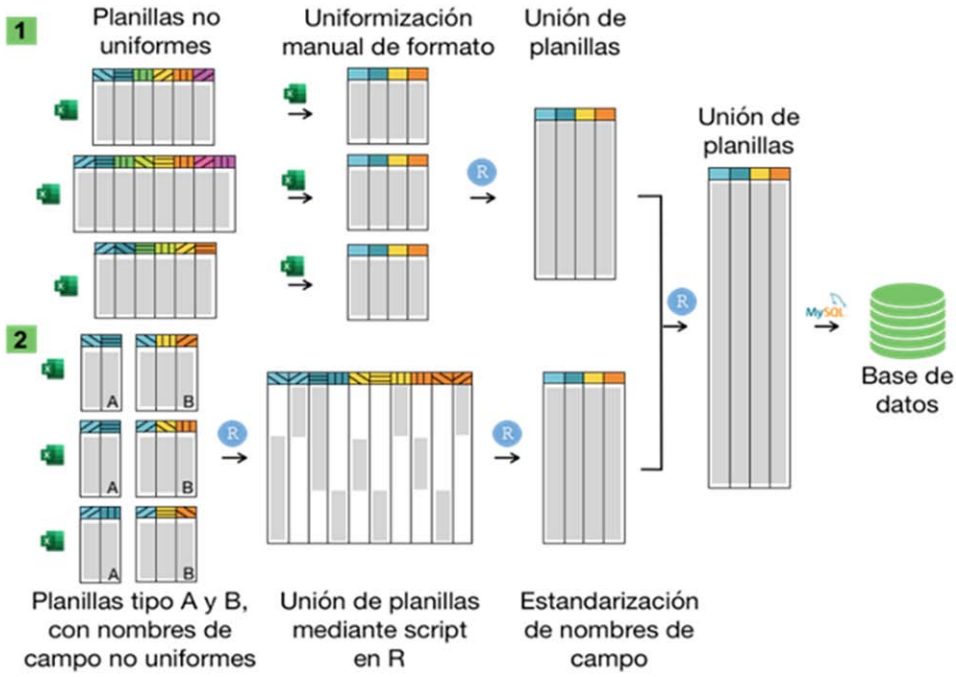

Figura 1. Esquema del proceso de consolidación de datos

RESULTADOS DE LA INVESTIGACIÓN

Se consolidaron un total de 2.047.947 *datapoints* agrupados en 87.175 registros correspondientes a 22 años, 948 ensayos y 15 localidades. El número de registros por año para cada subprograma fue de 41796 para *japónica* tropical, 21779 para *índica*, 5931 para Clearfield, 2098 para *japónica* templado, 1073 para Híbridos, y 8218 para las Evaluaciones Finales. En la figura 2 se muestra la distribución de los registros por año, localidad y etapa de evaluación.

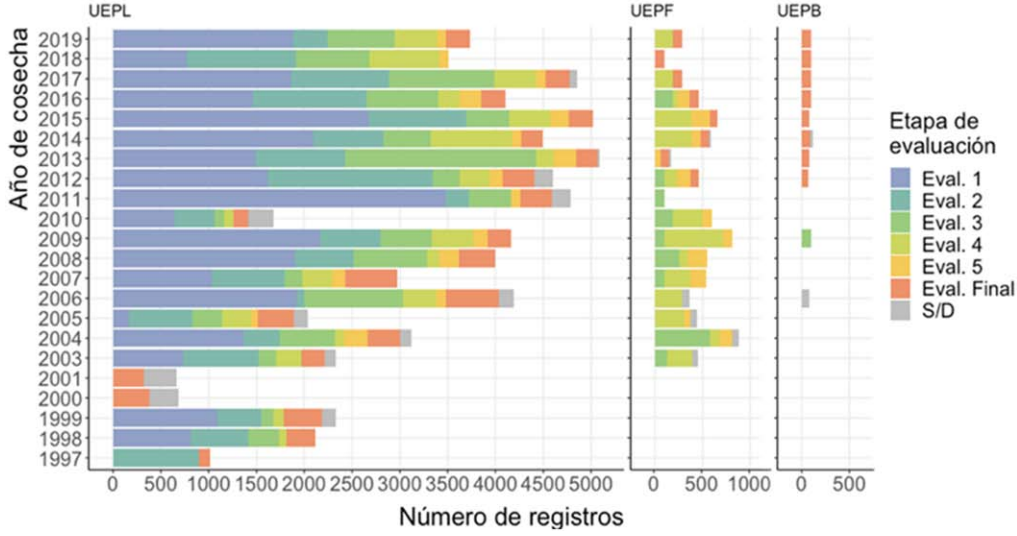

Figura 2. Cantidad de registros por etapa en cada localidad: Unidad Experimental Paso de la Laguna, Treinta y Tres (UEPL), Unidad Experimental Paso Farías, Artigas (UEPF) y Unidad Experimental Pueblo del Barro, Tacuarembó (UEPB).

En la figura 3 se muestra la conectividad entre años a través de líneas evaluadas, y de parentales cuyas progenies se evaluaron cada año. Se observa buena conectividad entre años, con tres genotipos (INIA Tacuarí, INIA Caraguatá y El Paso 144) evaluadas en común en los años más distantes (1997 y 2019), y con tres parentales cuyas progenies se evaluaron tanto en 1997 como en 2019 (El Paso 144, L-3616, y CL161). En general los años más cercanos están más conectados, tanto por líneas como por parentales, reflejando la repetición de líneas en las distintas etapas de evaluación del programa.

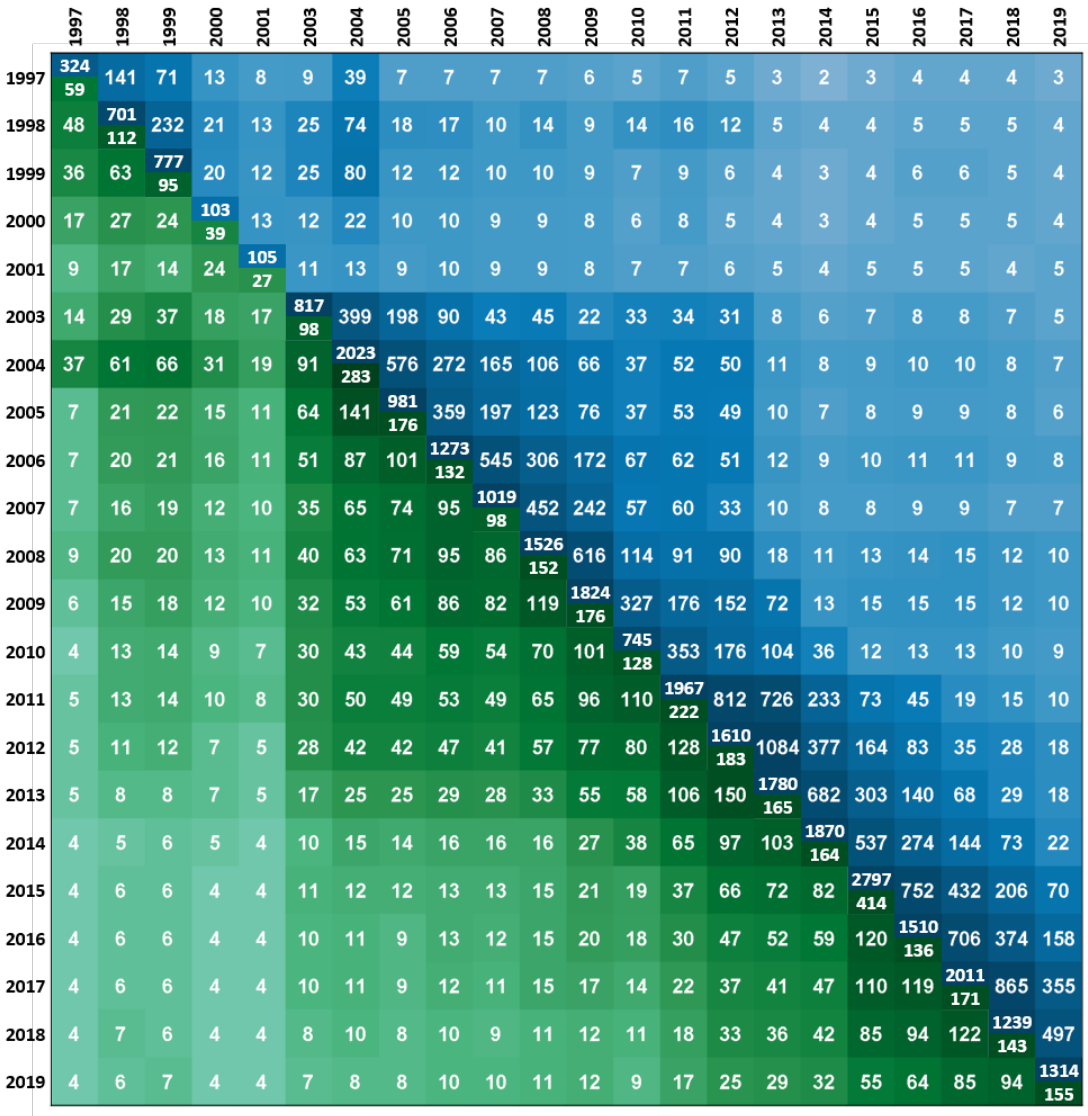

Figura 3. Conectividad entre años a través de líneas y parentales. En la diagonal se muestra en azul el número de líneas evaluado cada año, y en verde el número de padres cuyas progenies se evaluaron ese año. En el triángulo superior (azul) se muestra el número de líneas en común entre cada par de años, y en el triángulo inferior (verde) se muestra el número de parentales en común entre cada par de años.

La base de datos relacional implementada en SQL contiene cinco tablas (Líneas, Fenotipos, Genotipos, Cruzamientos, y Ensayos). La tabla Líneas contiene una identificación única para cada material genético evaluado por el PMGA, con posibilidad de registrar su identificación en el Banco de Germoplasma de INIA, así como la identificación autogenerada en programas de mejoramiento asociados como el de FLAR. Esta tabla se vincula a través de la identificación única con las tablas de Cruzamientos (información de pedigrí), Genotipos (información genotípica molecular), y Fenotipos (registros de variables medidas en ensayos de campo y laboratorio). Esta última se vincula con la tabla Ensayos a través de una identificación única para cada registro, que surge de la combinación automática del nombre del ensayo, el año, la localidad, la identificación de la línea, y la repetición.

CONCLUSIONES

A través de la creación la base de datos generada en este trabajo, se logró organizar la información generada por el PMGA de manera sistemática. Esto permitirá normalizar la información, concentrar todos los datos en un único lugar, evitar la redundancia y la duplicidad de registros, asegurar la integridad de la información almacenada y facilitar tanto el acceso como el registro de nuevos datos. Esta misma base de datos se continuará actualizando en forma rutinaria con los datos de cada año. Los datos actualmente consolidados facilitarán la realización de análisis conjuntos usando información de varios ensayos, lo que podrá en principio aumentar la precisión de las estimaciones y por ende la ganancia genética y la eficiencia del PMGA. También a partir de los datos consolidados se podrán hacer evaluaciones del progreso genético del PMGA a través de los años, así como hacer proyecciones sobre la ganancia genética futura.

BIBLIOGRAFÍA

Odell, S. G.; Lazo, G. R.; Woodhouse, M. R.; Hane, D. L.; Sen, T. Z. 2017. The art of curation at a biological database: principles and application. *Current Plant Biology*, 11: 2-11.
https:doi.org/10.1016/j.cpb.2017.11.001