Conjunto de datos de resultados de la Red Nacional de Cultivares de Girasol de INTA

Juan Caldera[1], Daniel Funaro[1], Lucio García[2], Julian Muñoz[2], Joaquin Seitz[2], Yanina Bellini Saibene[1]

Estación Experimental Agropecuaria Anguil "Ing.Agr. Guillermo Covas"
Ruta Nacional Nº 5, km 580, (6326) Anguil, La Pampa, Argentina
Colegio Universitario Liceo Informático II
Ameghino 865, (6300) Santa Rosa, La Pampa, Argentina
{caldera.juan, funaro.daniel, bellini.yanina}@inta.gob.ar

Resumen. Los resultados de evaluación de ensayos de cultivares contribuyen a la selección del híbrido que más se adapte al ambiente donde se realizará la siembra. Esta contribución presenta el armado y acceso de un conjunto de datos de 16 campañas agrícolas de evaluación de cultivares de girasol a nivel nacional realizados por la Red Nacional de Cultivares de Girasol de INTA. El conjunto de datos contiene 17.477 registros que corresponden a 897 Cultivares de 70 empresas evaluados en 331 ensayos realizados en 73. El conjunto de datos actual está disponible bajo licencia Creative Commons con Atribución y se pone a disposición de la comunidad académica, de productores y emprendedores.

Palabras claves: híbridos, elección cultivar, tecnología de cultivo, siembra

1 Introducción

El Instituto Nacional de Tecnología Agropecuaria (INTA), lleva adelante en todo el país una red de ensayos territoriales del cultivo de Girasol conocida como Red Nacional de Girasol de INTA (RNGINTA).

La elección del cultivar a sembrar se asocia a las características del cultivo: potencial de rendimiento, comportamiento sanitario, contenido de aceite, ciclo, altura y tipo de aquenio. Estas características determinan seguridad, productividad y rentabilidad del cultivo. El ambiente afecta el comportamiento de los cultivares en forma diferencial generando variaciones que son necesarias de interpretar y conocer [1].

La RNGINTA provee información que contribuye al conocimiento de la variabilidad genotípica, ambiental y de interacción Genotipo por Ambiente observada en los ensayos. Esta información, utilizada por asesores y productores, contribuye a la decisión de la selección del cultivar que mejor se adapte al ambiente donde se realizará la

siembra. Esta red se lleva adelante en convenio con la Asociación Argentina de Girasol (ASAGIR), a quienes se informan por localidad, los resultados finales de los ensayos realizados.

La información de la red se difunde por medio de un sitio web y cuadernillos impresos editados por INTA y ASAGIR, esto hace que para consultar o analizar información de varios años o varias localidades se deban descargar diferentes archivos. El armado de un único repositorio que presente la información de resultados de toda la red a nivel nacional permitiría simplificar el acceso a la información y realizar análisis temporales y espaciales a nivel de sitio, regiones, híbrido, empresa, por mencionar algunas dimensiones.

Este trabajo presenta un conjunto de datos unificado con más de 17.000 registros con los resultados de los ensayos de la RNGINTA de todo el país para las campañas 2001-2002 a 2016-2017.

Las actividades involucradas en este trabajo se realizaron en el marco de un convenio de comisión de estudios donde INTA trabaja en conjunto con instituciones educativas de nivel terciario y universitario del medio formando recursos humanos por medio de la resolución de problemas concretos del sector agropecuario. En este caso además orientado al desarrollo de información y herramientas digitales para la toma de decisiones de productores y asesores.

2 Contexto de la recopilación de datos

Como se explica en [1] y [2], la RNG INTA se integra con un conjunto de localidades y experimentos (las cantidades cambian con las campañas) donde profesionales del INTA y colaboradores son responsables de la elección de lotes para implantación de ensayos de cultivares, control de malezas y plagas, seguimiento, evaluación y toma de observaciones, recolección del material y procesamiento de los datos.

Los híbridos incluidos en cada ensayo son elegidos por los proveedores de semilla, quienes optan por aquellos que consideran aptos para ese ambiente. Participan entre 25 a 65 híbridos por localidad y año [2]. Cada uno de los ensayos cuenta a su vez con tres repeticiones si están dentro de una estación experimental agropecuaria (EEA) del INTA y cuatro repeticiones si están sembrados fuera de las EEAs [1].

ASAGIR publica en su sitio web [3] los resultados de cada campaña agrupado en tres regiones (Norte, Centro y Sur) y localidad (figura 1), esto resulta en un archivo diferente por cada campaña y localidad. El formato de los archivos de los datos es variable según campaña, pudiendo ser HTML, PDF o Excel; a su vez, los datos (columnas) presentes en cada planilla pueden variar de una localidad a otra o de una campaña a otra (figura 2).

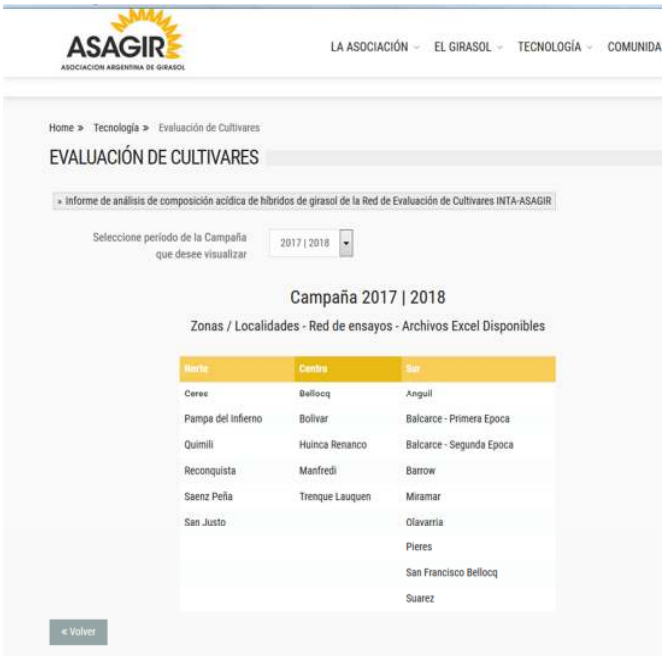

Fig. 1. Sitio web de descarga de los resultados de los ensayos de la RNGINTA

Para obtener un conjunto de datos unificado se llevaron adelante tareas de ordenamiento y limpieza de los datos. El concepto de ordenamiento de datos está relacionado a la estructura de los mismos mientras que el concepto de limpieza de datos está asociado al contenido.

El primer paso es generar una estructura ordenada para contener los datos de la red. Los datos están ordenados cuando cumplen tres criterios [4]:

1. Cada observación está en una fila.
2. Cada variable está en una columna.
3. Cada valor tiene su propia celda.

Cuando se realiza el ordenamiento de los datos, entonces se pueden limpiar de una forma más sencilla y eficiente.

El primer paso del proceso fue descargar todos los archivos disponibles en la web mencionada y transformar los archivos PDF a un formato de planilla de cálculo editable. Se utilizó Tabula [5], una herramienta gratuita para extraer datos de archivos PDF a archivos CSV y Excel, optándose por el formato xls. También se cambiaron a formato xls los archivos HTML. Este paso entregó como resultado **560 archivo**s editables en el mismo formato.

El siguiente paso consistió en realizar un análisis de los datos disponibles para diseñar una estructura relacional que cumpla con los requisitos de los datos ordenados. Este análisis detectó información en tres lugares de cada archivo:

4. El encabezado de la planilla: se detallan datos de la localidad del ensayo, la fecha de siembra, el responsable y su contacto (figura 2)
5. Las tablas de datos presentes en la planilla: contiene la información de los resultados, pero las variables informadas no son las mismas para todas las planillas (figura 2).
6. El nombre del archivo: se presentan datos de la localidad, el sistema de siembra utilizado o el tipo de ensayo (figura 3).

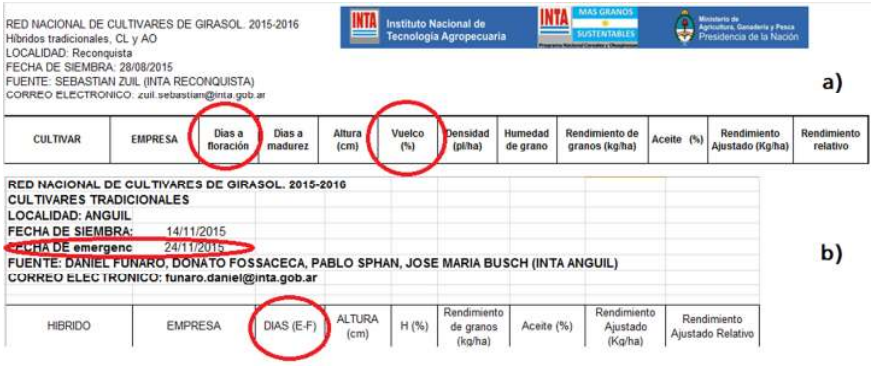

Fig. 2. Ejemplos de las planillas de la campaña 2015-2016 para dos localidades donde se aprecia que las columnas con datos difieren de una planilla a otra. Fuente: [3]

Este análisis arrojó como resultado un Diagrama de Entidad Relación (DER) que reflejó las entidades, las relaciones y todos los atributos que en algún momento se informaron (aunque no se informen en todas las planillas o en todas las localidades). El diseño ordenado incluye dos tareas más:

- Definir los tipos de datos para cada variable y la unidad de medida que le corresponde.
- Agregar variables y atributos no presentes en los datos, pero que permitan realizar un mejor aprovechamiento de los mismos.
- Agregar las claves de identificación única y las claves de relación entre entidades (claves primarias y claves foráneas)

El formato final del conjunto de datos con sus atributos, su tipo, unidad de medida y descripción se detallan en la tabla 1 y figura 4.

Campaña 2006 | 2007

Zonas / Localidades - Red de ensayos - Archivos Excel Disponibles

Norte	Centro	Sur
Caracterización sanitaria	Anguil	Ascasubi
Las Toscas	Anguil AO	Balcarce AO
Las Toscas AO	Bulnes	Balcarce CL
Las Toscas CL	Caracterización sanitaria	Balcarce fecha tard.
Reconquista	Depto. Gualeguay	Balcarce fecha temp.
Reconquista AO	Huinca Renancó	Barrow
Reconquista CL	Manfredi 52 Alta	Barrow AO
Roya Negra	Manfredi 70	Barrow CL
	Manfredi AO	Bellocq
	Manfredi CL	Belloq AO
	Paraná	Belloq CL
	Parana AO	C. Suarez Directa
	Paraná CL	Caracterización sanitaria

Fig. 3. Detalles del tipo de ensayo y sistema de siembra en el nombre del archivo: fecha tard., fecha temp, AO (Alto Oleico), CL (ClearField) y Directa indican tipo de ensayo, época y sistema de siembra utilizado. Fuente: [3]

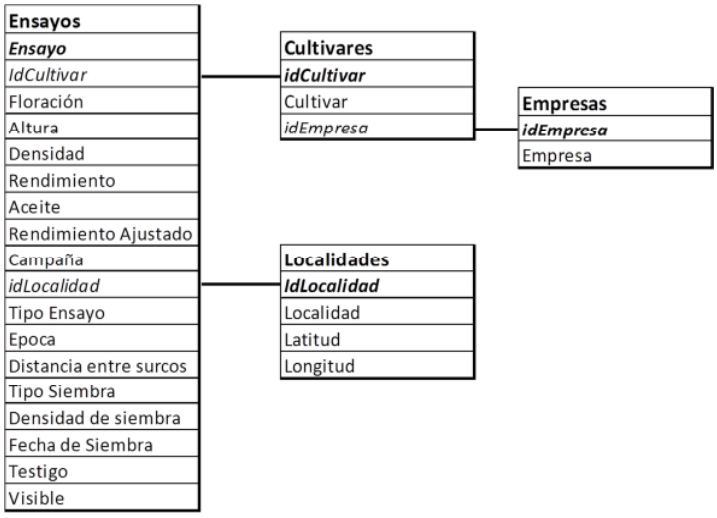

Fig. 4. Diagrama de Entidad Relación del conjunto de datos final para los resultados nacionales de la red de ensayos de girasol de INTA. Atributos en cursiva y negrita indican la clave primaria de la entidad, atributos en cursiva indican la clave foránea que representa la relación con otra entidad.

Tabla 1. Atributos, tipo de datos, unidad de medida y descripción del conjunto de datos final

Variable	Tipo	Unidad de Medida	Descripción
Cultivar	Carácter		Nombre comercial del híbrido [2]
Empresa	Carácter		Semillero proveedor del híbrido [2] (obtentor actual)
Floración	Entero	días	Días entre siembra y floración [2]
Altura	Entero	cm	Altura promedio de plantas de las parcelas, medida en floración [2].
Densidad	Entero	pl/ha	Densidad promedio de plantas de las parcelas evaluadas, expresada en plantas por hectáreas
Rendimiento	Entero	kg/ha	Rendimiento de aquenio expresado en kg/ha, a 11% de humedad[2]
Aceite	Decimal	%	Contenido porcentual de aceite del aquenio medido por Resonancia Magnética Nuclear o RMN. Expresado en base seca. [2]
Rendimiento Ajustado	Entero	kg/ha	Rendimiento calculado a través de una fórmula que combina el rendimiento de aquenio y el contenido de aceite, transformando en kg/ha la bonificación o descuento (2% por cada punto con respecto a la base) que corresponde al porcentaje de aceite con base de comercialización de 42%. El valor obtenido se suma (bonificación) o resta (descuento) al rendimiento obtenido en kg/ha.[2]
Campaña	Carácter		Años de la campaña. Inicial: 2001-2002
Localidad	Carácter		Nombre de la localidad donde se realiza el ensayo
Latitud	Decimal		Coordenadas de la localidad donde se realiza el ensayo
Longitud	Decimal		Coordenadas de la localidad donde se realiza el ensayo
Ensayo	Carácter		Nombre del ensayo que se forma con la localidad y la campaña
Tipo Ensayo	Carácter		Puede ser: Convencional, Clearfield, Alto Oleico, Confitero, Tardia
Epoca	Carácter		Puede ser: Normal, Segunda, Tardia, Temprana
Distancia entre surcos	Entero	cm	Distancia entre surcos en centímetros
Tipo Siembra	Carácter		Puede ser: Directa, Convencional
Densidad de	Entero	semillas	Densidad de siembra

siembra			
Fecha de Siembra	Fecha		Fecha en que se sembró el ensayo
Testigo	Lógico		Testigo del ensayo
Visible	Lógico		Si se presenta el dato o no par a algunas salidas

Con la estructura de datos ordenados, el tercer paso fue unificar los 560 archivos descargados en un solo conjunto de datos respetando la estructura diseñada (tabla 1). Luego se inició la limpieza de los datos analizando el contenido de cada columna con el software OpenRefine [6]. Este software es una herramienta para poder limpiar datos y transformarlos en diferentes formatos. Los problemas resueltos con OpenRefine fueron:

1. Mismo cultivar escrito de diferente manera o con errores en el nombre.
2. Misma empresa escrita de diferente manera o con errores en el nombre.
3. Diferentes empresas para un mismo cultivar. Se unificó colocando el nombre de la empresa que actualmente ostenta el cultivar.
4. Localidades escritas de diferente manera o con errores en el nombre.
5. Transformación de los valores de aquellos atributos que se encontraban en unidades de medidas diferentes a la medida seleccionada.
6. Completar datos faltantes.

Finalmente se completaron datos como la latitud y longitud de la localidad del ensayo. En caso de contar con las coordenadas del lugar del ensayo (figuraba en el encabezado de algunas planillas) se utilizaron estas coordenadas, de lo contrario se colocaron las coordenadas de la localidad indicada obtenidas con Google maps.

Para poder unificar futuros archivos de resultados de la red, se almacenó el nombre original con el que figuraba el cultivar en los 560 archivos, junto con el nombre correcto del cultivar. Esta acción permite reconstruir los datos de la fuente original y asignar el nombre correcto del cultivar en caso que se encuentre con algún error ya conocido en el procesamiento realizado.

Además OpenRefine lleva un historial de cambios realizados sobre la fuente de datos, por lo que los pasos realizados están registrados en un archivo y se pueden deshacer y replicar.

3 Resultados

El trabajo presentado ordenó y limpió 560 archivos en un solo conjunto de datos que contiene 17.477 registros que corresponden a 897 Cultivares de 70 empresas evaluados en 331 ensayos realizados en 73 localidades durante 16 Campañas.

El listado obtenido de cultivares y empresas se está utilizando en el sistema de información de gestión de la Red Nacional de Cultivares de Girasol que se está desarrollando en la EEA Anguil [1]. Este listado unificado permite que los errores de tipeo, formas diferentes de escribir un híbrido, cambios de empresas, etc, sean transparentes al sistema, al seleccionar los cultivares desde una lista única para todo el país. Igualmente redunda en la comunicación de los resultados, al presentar siempre el mismo dato para todas las localidades.

Además, el conjunto de datos permite analizar geográfica y temporalmente los resultados de los ensayos de la red, en las figuras 5 a 7 se presentan una serie de ejemplos de resultados del análisis del conjunto de datos completo.

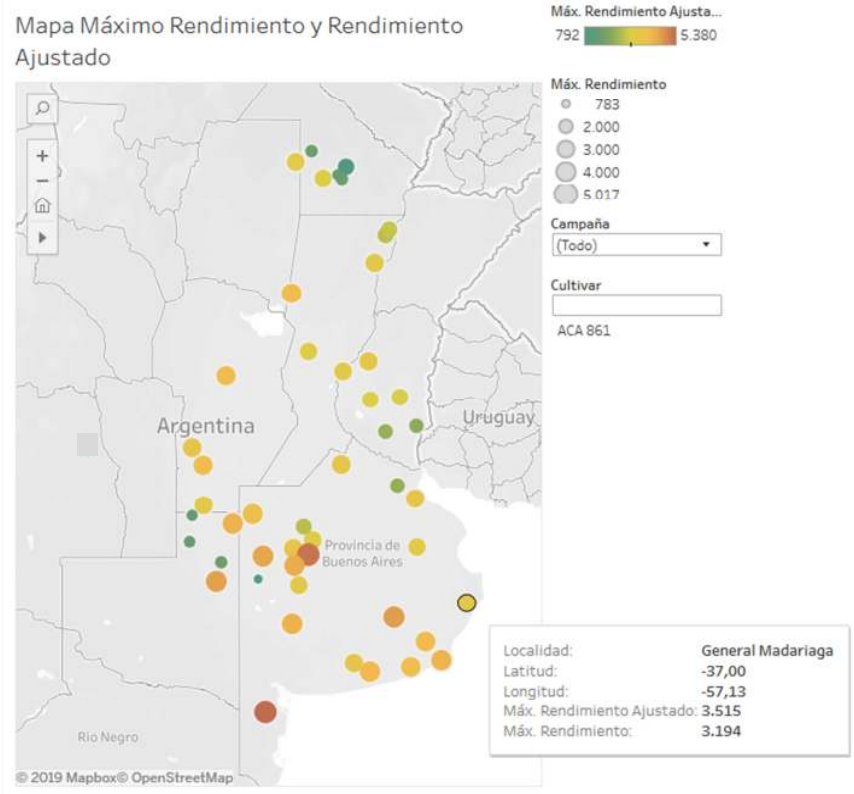

Fig. 5. Consulta espacial de máximo rendimiento ajustado. Filtros: Cultivar ACA 861 en todas las campañas.

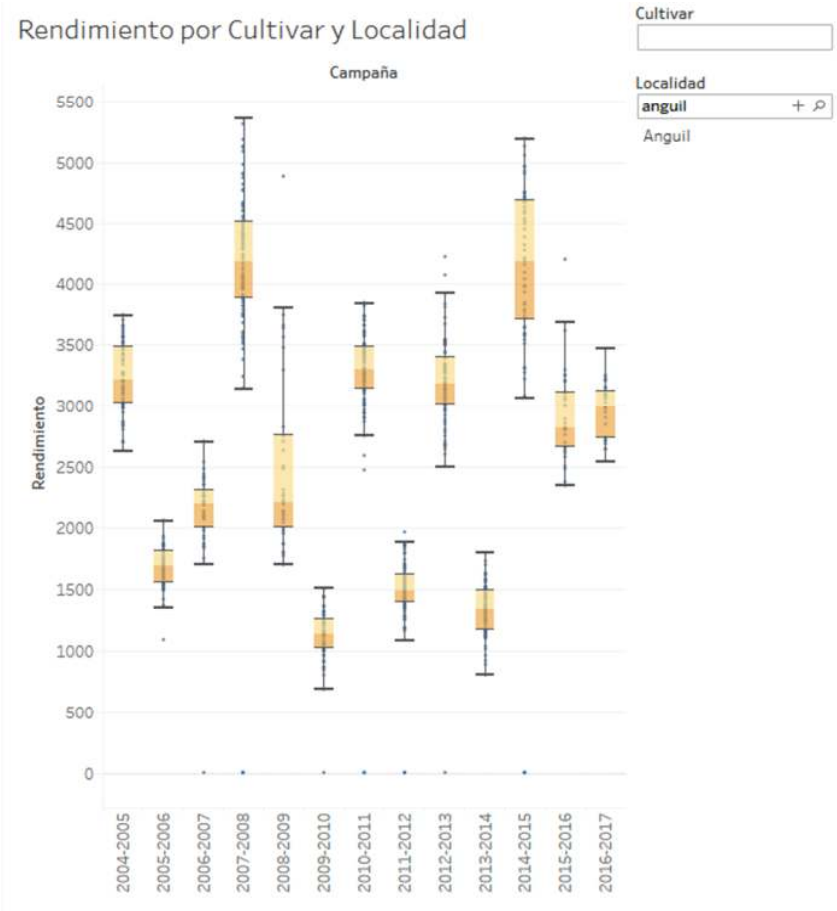

Fig. 6. Rendimiento por cultivar y localidad. Filtros: todos los cultivares evaluados, en todas las campañas para la localidad de Anguil.

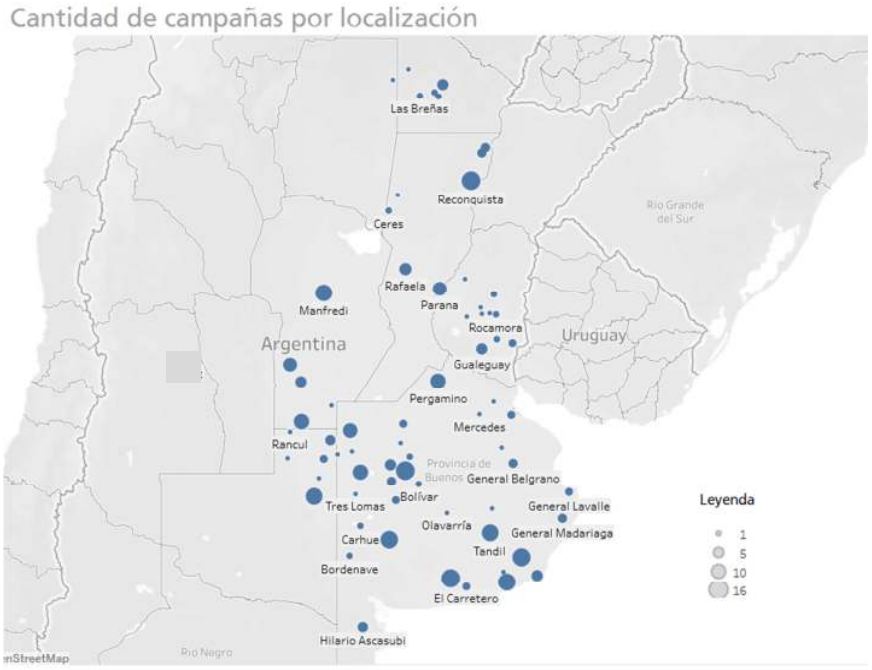

Fig. 7. Cantidad de campañas en las que se realizaron ensayos en cada localización en todo el período estudiado.

Estas figuras de ejemplo presentan información de interés para la toma de decisiones productivas, en la elección del cultivar a utilizar, como las figuras 5 y 6, pero también se puede utilizar para decisiones de gestión de la RNGINTA, como por ejemplo la figura 7 donde se visualiza la cobertura de la red y la cantidad de campañas en la que participa cada localización.

Este set de datos se continuará actualizando las sucesivas campañas que se publiquen en la web de ASAGIR. Se entrega el mismo con una con una Licencia Creative Commons Atribución 4.0 Internacional, según la cual, en cualquier explotación que se haga de los datos será necesario reconocer la autoría de forma obligatoria [7]. El mismo se encuentra disponible a pedido al mail caldera.juan@inta.gob.ar. Se están realizando los trámites para su publicación en el repositorio institucional.

Como trabajos futuros se ha planificado la generación de un paquete de R para que contenga los datos y presente una serie de funciones que ayuden a la elección del cultivar.

4 Referencias

[1] J. Caldera y D. Funaro, «Propuesta de modelo de datos para la Red Nacional de Evaluación de Cultivares Comerciales de Girasol», en *VIII Congreso Argentino de AgroInformática (CAI-2016)-JAIIO 45 (Tres de Febrero, 2016)*., 2016.

[2] «Red Nacional de Evaluación de Cultivares Comerciales de Girasol del INTA. Region centro. CICLOS 2006-2007 / 2007-2008», INTA - ASAGIR, Buenos Aires, Argentina, Cuadernillo informativo 13, ago. 2008.

[3] «Evaluación de Cultivares», *ASAGIR - Asociación Argentina de Girasol*. [En línea]. Disponible en: http://www.asagir.org.ar/acerca-de-evaluacion-de-cultivares-463. [Accedido: 02-may-2019].

[4] H. Wickham, «Tidy Data», *J. Stat. Softw.*, vol. 59, n.º 1, pp. 1-23, sep. 2014.

[5] «Tabula: Extract Tables from PDFs», 16-jun-2013. [En línea]. Disponible en: http://tabula.technology/. [Accedido: 02-may-2019].

[6] Google, «http://openrefine.org/», *OpenRefine*. [En línea]. Disponible en: http://openrefine.org/. [Accedido: 07-may-2019].

[7] «Creative Commons — Attribution 4.0 International — CC BY 4.0». [En línea]. Disponible en: https://creativecommons.org/licenses/by/4.0/legalcode. [Accedido: 22-abr-2019].