Scientific quarterly journal ISNN 1429-7264

Agricultural Engineering

2014: 3(151):227-235

Homepage: http://ir.ptir.org

DOI: http://dx.medra.org/10.14654/ir.2014.151.075

UPDATING DATABASES OF THE INTERNET DECISION SUPPORT SYSTEM FOR CEREALS PROTECTION[1]

Andrzej S. Zaliwski*, Anna Nieróbca

Department of Agrometeorology and Applied Informatics
Institute of Soil Science and Plant Cultivation – State Research Institute in Puławy
*Contact details: ul. Czartoryskich 8, 24-100 Puławy, e-mail: andrzej.zaliwski@iung.pulawy.pl

ARTICLE INFO

Article history:
Received: August 2014
Received in the revised form: September 2014
Accepted: September 2014

Keywords:
data quality
information quality
data updating, database
decision support system

ABSTRACT

The process of data updating in the databases of the Internet decision support system for cereals protection is presented. The system uses three kinds of data: on spring and winter wheat varieties, on plant protection products (fungicides and insecticides against wheat diseases and pests) and on product doses. It was found that it is possible to obtain all the necessary data for the updating from available web resources. Data sources for variety information are the COBORU materials and websites of agricultural plant breeders. The search engine of plant protection products available on the website of the Ministry of Agriculture and Rural Development is an essential source of the data on protection products and their doses. Other sources of these data are websites of plant protection product producers and online shops. In the data updating process the following stages were distinguished: owned-resources stocktaking, data quality criteria selection, credible data sources determination, data acquisition by the application of the criteria, data adaptation to the system requirements, data entry, data verification. Each one of the itemized stages contributes to the ultimate data quality after the updating. The stages of credible data sources determination, data acquisition and data adaptation to the system requirements were found to be labour intensive. About a triple reduction of labour input into the updating was achieved through the improvement of the data acquisition process, gaining proficiency in the database editors operation and the cumulative effect of build-up of specific data resources in the system. The adopted verification procedures for update correctness are described.

Introduction

"The Internet decision support system in cereals protection" (CP-DSS) constitutes an element of "The decision support system in the integrated plant protection" (Nieróbca et al., 2010; Zaliwski, 2009). The mentioned supersystem, being still in the prototype stadium was marked with the acronym IPM DSS (*Integrated Pest Management Decision Support*

[1] Publication written as part of project 4.1. of the multi-annual programme of IUNG-PIB Puławy.

System). IPM DSS was created as a result of realization of the foreign research-development project, carried out together with the Danish Institute of Agricultural Sciences. The main elements of IPM DSS are plant protection models generating recommendations on the need to carry out treatments based on precisely determined threshold values (Nieróbca, 2009). Economic thresholds are determined with regard to the percentage of plants with disease symptoms. Exceeding the threshold suggests the increase of the yield losses above the treatment costs which justifies the protection treatment. Both economic thresholds and doses of plant protection products, recommended by the system, are defined for the specific species, variety and growth stage.

The developed economic thresholds were verified in field experiments in various environmental conditions (Nieróbca, 2009). Field experiments carried out in 2001-2006 confirmed the usefulness, in Polish conditions, of winter and spring wheat protection models against the most important diseases: mildew of grasses and cereals, yellow rust, brown rust, eyespot and septoria as well as against pests: aphids and grain beetle larvae.

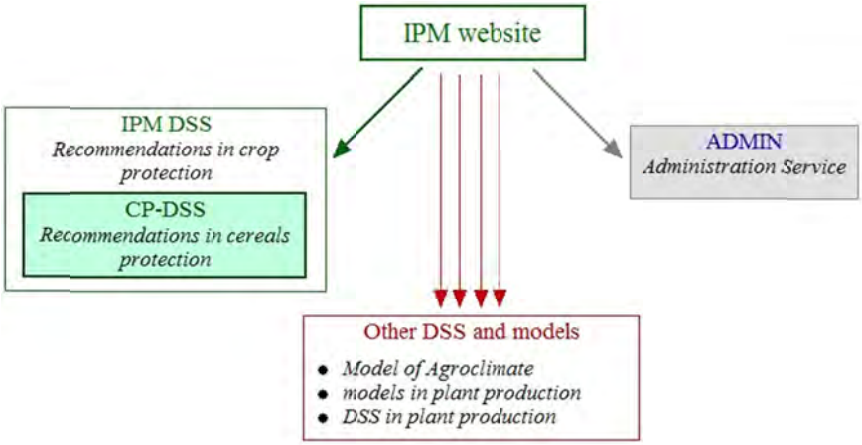

Figure 1. Organization of the IPM website

IPM DSS system was made available in the Internet in 2003 as an element of the website marked with the IPM acronym (fig. 1). PC back office programs for management of data bases are an important addition to the system, which was provided to the Polish team by the Danish party. They enabled a manual edition of the data on varieties, plant protection products, users and weather data. With their use, update of data was carried out to the moment of the IPM website migration to a new server in 2007. The change of the software environment (operational system, database) immobilized Danish back office programs, disabling data updating. So, soon a decision was made to develop own back office software (Zaliwski, 2009b; Zaliwski, 2014). Frameworks of the internet "Administration Service" (fig.1) were designed, designated for management of not only the IPM DSS databases but of all the internet decision support systems and models developed in the Department of Agrometeorology and Applied Informatics of the IUNG-PIB (Zaliwski, 2009a; Zaliwski and Nieróbca, 2013). The ambitiously envisioned programme of development of the

Administration Service, in comparison to the work team potential, resulted in a considerably late construction of the back office applications patterned after the Danish PC programs, held up till 2013-2014. These works were carried out simultaneously with the second migration of the IPM DSS system to the next server (Zaliwski and Nieróbca, 2013), which enabled removal of unnecessary data "inherited" from the Danish prototype.

When starting data update in a big system, a specific technology for this process should be assumed. Thorough update carried out for the first time requires preparation of data, tools for data edition and development of proper procedures.

The objective of the article is the presentation of the databases update process of the CP-DSS carried out in 2013-2014.

Preparation of database update

The main principle in the process of data entry to the system is a care for their high quality in the process of collection. Later verification and correction, after the entry has been conducted, is labour intensive, expensive and ineffective. Experts on the data and information quality, such as Larry P. English (1999; 2009) and Thomas C. Redman and A. Blanton Godfrey (1997) take the same view.

The first stage of data update is an inventory of the resources owned in order to know their structure and in order to find out the lacking elements. Such an inventory was quite easy to conduct with the use of the back office applications for edition of databases of the CP-DSS (Zaliwski, 2014). Inventory results (in a rough outline) are presented in figure 2.

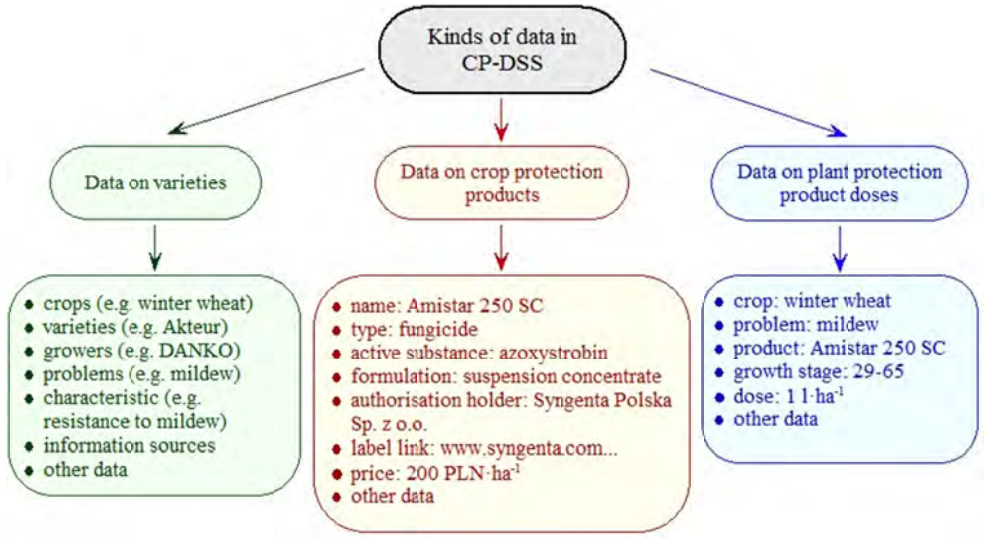

Figure 2. Kinds of data in Cereal Protection DSS

In the second stage credible data sources should be chosen. With reference to credibility Józef Oleński (2001) divides information sources into original, secondary and derivated. Original sources of information are real objects, phenomena and processes. Secondary

information sources are organizations (institutions) which use their own information systems for generation of information based on primary sources. Derivated information comes directly or indirectly from secondary sources. The bigger the number of intermediate stages, the further copy of the original we deal with and the more errors and inaccuracy may occur.

An example of the secondary source of data on varieties is "*The list of agricultural plants entered into the national register in Poland*" [Lista odmian roślin rolniczych wpisanych do krajowego rejestru w Polsce] (2013) and websites of crop variety growers. For data on plant protection products these are: "*Search engine of crop protection products*" [*Wyszukiwarka środków ochrony roślin*] (2014), which can be accessed on the website of the Ministry of Agriculture and Rural Development as well as labels and material safety data sheets of plant protection chemicals. Labels and safety data sheets are obtained from websites of producers or distributors of products (those who hold an authorisation to introduce a product on the territory of Poland). Online shops are an example of the secondary source of crop protection product prices. An online shop is not a relevant data source concerning characteristics of the crop protection product because it provides derivated data.

Data acquisition is the next stage. This stage is quite complex. One should compare the offer with the data quality criteria, such as the criterion of compatibility with the requirements of the system, completeness, timeliness and correctness. The compatibility criterion means that only those data are entered which will be used by the system. In the CP-DSS, the plants for which protection models have been implemented are determinants in this regard. Presently, there are only models for spring and winter wheat, which include the pests mentioned in the introduction. Meeting the conditions of completeness and timeliness means a necessity to omit the records, particular parts of which are incomplete and not up-to-date.

With reference to completeness a criterion "stiff" or "soft" may be assumed. The "stiff'" criterion means a necessity to fill the entire record without leaving anything out. The "soft" criterion differentiates between two parts of the record: an identification and calculation part and an informative part. Requirement of completeness concerns only those record fields, which are used for generation of recommendations (identification and calculation part). In the remaining fields (informative part) lack of data is admissible. For example, a crop protection product must have a name and type (fig. 2) because these are identification fields. Price and dose are also indispensable, because they are used for calculation of the treatment cost. However, many other data on the product, such as toxicity, threat to bees, optimal conditions for effectiveness, impact on the successive crops or the risk of immunization constitutes an additional piece of information, less or more significant to the user. The use of the soft completeness criterion of a varied degree of rigour makes it possible to omit specific data of informative nature. It is important from the practical point of view, because a full set of data on the plant protection product in the databases of the CP-DSS constitutes a set of 44 various attributes. Some data are difficult to acquire and they are less important to the user, e.g. the risk of immunization of the pest against an active substance. The use of the stiff criterion would foreclose the use of such a record, limiting in practice the update on account of numerous deficiencies in available data.

The criterion of data timeliness in the case of varieties takes into account the expiration date of the entry in the national register (Lista odmian roślin…, 2013). In the case of crop protection products, the date of product approval expiration is such an indicator (Wyszuki-

warka środków…, 2014). The search engine of crop protection products of the Ministry of Agirculture and Rural Development enables searching the ministry database on crop protection products according to different attributes. It enabled the construction of the list of fungicides and insecticides against diseases and pests of winter and spring wheat with the date of approval expiration later than 2014. The list enabled a systematic approach to updating of the databases on crop protection products of the CP-DSS (Zaliwski and Nieróbca, 2014) eliminating at the beginning the products soon to be withdrawn from use (soon to reach the expiration date) and those which are not useful for the system on account of the pests controlled.

Specific data, such as growth stages, most frequently referred to in source materials by name, required to be expressed in numerical scale. In this case a table from Wikipedia (Skale BBCH, 2014) was used. Expression of variety resistance to pests in the scale used in the CP-DSS was a more difficult task. Original data on resistance were obtained from the website of COBORU "*Post-registration variety testing and recommendation*" [Porejestrowe doświadczalnictwo odmianowe i rolnicze] (PDOiR, 2014), and then they required to be interpreted by an expert.

Data entry

Sets of data on varieties, crop protection products and doses were introduced to the databases of the CP-DSS manually with the use of three various back office applications (editors): "editvariety", "editchemicals" and "editdiseasepest" (Zaliwski, 2014). The most time-consuming was the entry of the data on crop protection products on account of extensiveness of the data set. In the initial period of update, the entry of one set of data took approx. 4 hours. Searching for a data source in the Internet consumed much time at the beginning (finding websites of producers and distributors of products, finding online shops).

The procedure of searching for data on crop protection products is as follows. After drawing a product from the list of products (prepared earlier pursuant to the expiration date and the pests controlled) its availability on the market is checked. Its price and the size of packages should be found in internet shops. Without these data, the product may be discarded outright as non-useful for the CP-DSS. In the end, a label and a safety data sheet is downloaded from the website of the producer or the distributor. After the collection of these data, the crop protection product may be entered into the database. The data are copied into the editor straight from the WWW pages as well as from labels and data safety sheets (available by default as PDF files). In case of non-editable PDF files they were converted to editable files.

It was found out that the main sources of significant work input of manual updating of the data on crop protection products were:

- large number of attributes (44),
- no strict standardization of labels and data safety sheets (difficulty in finding data),
- no rigour concerning the content of provided information (necessity of converting information to single format),
- difficulty in determination of the URL address of labels and data safety sheets or no data safety sheets,
- necessity of shortening the content in order to adjust the length to some fields in database,

– necessity to navigate through many pages of the editor (application "editchemcials" has e.g. 10 pages).

This last source of labour intensity of manual data update was the incentive for the attempt to automatize data entry with the use of „ipmDataLoader" program (Zaliwski, 2014). The program enables import to the databases of the CP-DSS entire variety and crop protection product data sets in one go. It was successfully used for import of data on varieties, whereas in the case of plant protection products the attempt was not successful. Before import the data require preparation in the text file, which took almost the same time as their copying into the editor "editchemicals".

Becoming skilled at searching for data and working with the application "editchemicals" allowed after some time approx. threefold increase in efficiency of product data entry. Of importance was an accumulative effect of build-up of specific data, which once entered, were available in the selection boxes.

Verification of correctness of the data entered

It should be noticed that the databases of the CP-DSS by themselves have basic control mechanisms of the correctness of the data being entered (e.g. minimum and maximum for numerical values, dates, etc.) However, a great number of data are in a text format. Construction of validation rules in this case would be very difficult on account of unpredictability of content. Thus, validation carried out by humans is required. Then, however, even the highest concentration and care do not guarantee complete elimination of errors during data input.

Figure 3. WWW page displaying characteristic of protection chemical

Data verification was carried out in two ways. First of all, just after the input, the input version was compared with the source. The second way was checking the manner of the input data display in the front office applications of the CP-DSS:

– on the recommendation-generating page,
– on the page displaying the chemical product characteristics.

Verification of data in the front office application allows finding out and correcting errors made during their entry and overlooked during the first verification. It requires comparing two pages. In the front office application, the data are verified and in the database editor the errors are corrected. Such work should be carried out at the workstation equipped with two screens. An example of finding out the missing datum "Formulation type" (marked with a red frame) in the application "Characteristics of protection chemical" of the CP-DSS system is presented in figure 3.

Summary and conclusions

The tasks which were carried out, with the aim to update the databases of the CP-DSS system make it possible to draw the following conclusions:

1. Generally available internet resources were a sufficient source for obtaining all necessary data for the CP-DSS update. Some data (e.g. plant resistance) required expert interpretation in order to adjust to the scale used in the system.
2. On account of the manner of obtaining the data (independent acquisition from the sources available in the Internet), the update was time-consuming at the beginning, which was caused by the necessity to find information sources and acquiring skills at searching for the necessary resources.
3. Main reasons of significant work input in the process of updating data on protection chemicals was the difficulty in determination of URL addresses of some products (non-intuitive construction of WWW pages), difficulty in finding data in the texts acquired (inconsequential use of the content standardization by information generators) as well as the necessity to shorten the data content (in adjustment to the restrictions of the CP-DSS databases).
4. Improvement of the process of data acquisition and their adjustment to the system needs, acquiring skills at editors operation and accumulative effect of specific data build-up in the system had great impact (approx. three times) on the reduction of labour input of the update.

The following stages of the update process of databases were distinguished in this paper: inventory of the resources owned, selection of credible data sources, selection of the data quality criteria, data acquisition with the use of the criteria, adjusting data to the system needs, verification of the data entered. Each of the mentioned stages affects the final (after-update) quality of data. When assessing the tasks carried out, one should bear in mind that after every effort has been made, errors may occur, which is proved by figure 3. However, working out a reliable update process has a great impact on their reduction. Thus, data update is a process, which should be constantly improved.

References

English, L.P. (1999). *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. New York, John Wiley & Sons. ISBN: 978-0-471-25383-9.

English, L.P. (2009). *Information Quality Applied: Best Practices for Improving Business Information, Processes, and Systems*. New York, John Wiley & Sons. ISBN: 978-0-470-13447-4.

Lista odmian roślin rolniczych wpisanych do krajowego rejestru w Polsce. Słupia Wielka, COBORU, 2013. Obtained from: http://www.coboru.pl.

Nieróbca, A. (2009). *Systemy wspomagania decyzji w ochronie roślin jako element integrowanej produkcji*. Studia i Raporty IUNG-PIB, 16, 31-44.

Nieróbca, A., Zaliwski, A. S., Horoszkiewicz-Janka, J. (2010). Rozwój internetowego systemu wspomagania decyzji w ochronie zbóż. *Inżynieria Rolnicza, 7*(125), 167-173.

Oleński, J. (2001). *Ekonomika informacji: podstawy*. Warszawa, PWE. ISBN 83-208-1311-5.

Pesticide Properties DataBase. Hatfield, Hertfordshire, UK, Agriculture & Environment Research Unit, University of Hertfordshire, (2013). Obtained from: http://sitem.herts.ac.uk/aeru/footprint/pl/index.htm.

PDOiR. *Porejestrowe doświadczalnictwo odmianowe i rolnicze*. Słupia Wielka, COBORU, (2014). Obtained from: http://www.coboru.pl/DR/index.aspx.

Redman, T. C., Godfrey, A. B. (1997). *Data Quality for the Information Age*. Artech House Publishers. ISBN: 978-0890068830.

Skale BBCH. Wikipedia, 2014. Obtained from: http://pl.wikipedia.org/wiki/Skale_BBCH.

Wyszukiwarka środków ochrony roślin. Warszawa, MRiRW, 2014. Obtained from: www.minrol.gov.pl/pol/Informacje-branzowe/Wyszukiwarka-srodkow-ochrony-roslin.

Zaliwski, A. S. (2009a). Ogólna koncepcja krajowego systemu wspomagania decyzji w zakresie produkcji roślinnej. *Inżynieria Rolnicza, 6*(115), 323-329.

Zaliwski, A. S. (2009b). *Organizacja modułu pogodowego krajowego systemu doradztwa w zakresie zrównoważonej produkcji roślinnej.* Studia i raporty IUNG-PIB, 16, 107-117.

Zaliwski, A. S. (2014). *Oprogramowanie narzędziowe portalu IPO*. Studia i raporty IUNG-PIB, 38(12), 89-114.

Zaliwski, A. S., Nieróbca, A. (2013). *Migracja portalu IPM do Windows Server 2008*. Studia i raporty IUNG-PIB, 33(7), 79-95.

Zaliwski, A. S., Nieróbca, A. (2014). *Aktualizacja danych systemu IPM DSS*. Studia i raporty IUNG-PIB, 38(12), 67-87.

AKTUALIZACJA BAZ DANYCH INTERNETOWEGO SYSTEMU WSPOMAGANIA DECYZJI W OCHRONIE ZBÓŻ

Streszczenie. Przedstawiono proces aktualizacji danych w bazach danych internetowego systemu wspomagania decyzji w ochronie zbóż. System wykorzystuje trzy rodzaje danych: o odmianach pszenicy jarej i ozimej, o środkach ochrony roślin (fungicydy i insektycydy przeciw chorobom i szkodnikom pszenicy) i o dawkach środków. Stwierdzono, że wszystkie niezbędne dane do aktualizacji można pozyskać z dostępnych zasobów internetowych. Źródłem danych o odmianach są materiały COBORU oraz strony internetowe hodowców roślin rolniczych. Istotnym źródłem danych o środkach ochrony roślin i ich dawkach jest wyszukiwarka środków ochrony roślin udostępniana na stronach Ministerstwa Rolnictwa i Rozwoju Wsi. Inne źródła tych danych to strony internetowe producentów środków ochrony roślin oraz sklepy internetowe. W procesie aktualizacji danych wyróżniono następujące etapy: inwentaryzacja posiadanych zasobów, wybór kryteriów jakości danych, ustalenie wiarygodnych źródeł, pozyskanie danych z zastosowaniem kryteriów, dostosowanie danych do potrzeb systemu, wprowadzanie, weryfikacja danych. Każdy z wymienionych etapów ma wpływ na ostateczną jakość danych po aktualizacji. Stwierdzono dużą pracochłonność etapów: ustalenia wiarygodnych źródeł, pozyskiwania danych i dostosowania danych do potrzeb systemu. Około trzykrotne zmniejszenie pracochłonności aktualizacji uzyskano na skutek ulepszenia procesu technologicznego pozyskiwania danych, zdobycie doświadczenia w obsłudze edytorów baz danych i kumulatywny efekt wzrostu zasobów określonych danych w systemie. Opisano przyjęte procedury weryfikacji poprawności wprowadzenia danych.

Słowa kluczowe: jakość danych, jakość informacji, aktualizacja danych, baza danych, system wspomagania decyzji