

<https://doi.org/10.62405/osiconf.2026.53>

Zendulková Danica,

Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia
ORCID: 0000-0003-2487-0177, e-mail: Danica.zendulkova@cvtisr.sk

SOFTWARE TOOLS FOR DATA QUALITY MANAGEMENT IN THE CRIS SYSTEM

ABSTRACT. *The Slovak Current Research Information System (SK CRIS) integrates research information from multiple heterogeneous sources, which creates significant challenges for ensuring data completeness, accuracy and consistency. This paper summarizes practical experience with data quality management gained during the operation of SK CRIS and presents software-supported approaches to improving research information quality. Particular attention is paid to research projects, including batch data import, researcher–project matching, identification of incomplete and duplicate records, and data enrichment. The implemented solutions combine built-in validation mechanisms, API-based processing using Postman, batch processing of XLSX data and controlled vocabularies. It also explores the potential of artificial intelligence for assigning missing research fields based on project titles and abstracts. The results demonstrate that effective CRIS data quality management requires a combination of automated validation, data processing and enrichment tools with expert supervision and continuous monitoring.*

KEYWORDS: *Current Research Information System; Data quality management; Data cleaning; Data monitoring; SK CRIS; Software tools.*

INTRODUCTION. The Slovak Current Research Information System SK CRIS is a public administration information system and is managed by the Ministry of Education, Research, Development and Youth of the Slovak Republic pursuant to Act 172/2005 Coll. on State Support for Research and Development. It represents one of the basic pillars of the state science policy (Zendulková et al., 2026). The operation of the system is entrusted to the Slovak Centre of Scientific and Technical Information.

The SK CRIS system by law contains structured information on science and research entities (organizations and researchers), their activities (research projects) and results (mainly publications). Data are obtained from the information systems of grant agencies¹, from individual entities, including data provided within the framework of legally defined processes and data collections, as well as through system administrators.

The integration of research information from heterogeneous sources is a common data quality challenge in CRIS systems (Azeroual & Schöpfel, 2019). The heterogeneity of sources and the different level of validation of input data represent a significant challenge in terms of ensuring their quality. For this reason, it is necessary to create a systematic and effective data quality management system. This problem primarily concerns records of research projects, but to a lesser extent also other objects registered in SK CRIS.

Data quality management cannot be based only on a high-quality administrator interface. It also requires software tools that allow preventing duplication, identifying incomplete, inconsistent or otherwise problematic records and supporting the completion of missing data.

The aim of the article is to summarize the main problems and practical experience in the field of data quality management gained during the operation of SK CRIS, to present selected software tools and procedures used to improve data quality and to outline the possibilities for their further development.

¹ The Slovak Research and Development Agency (APVV), the Scientific Grant Agency (VEGA), the Cultural and Educational Grant Agency (KEGA), Research and Innovation Authority (VAIA)

Current situation

The SK CRIS², information system, built in accordance with European standards and based on the CERIF³ relational data model, has been in operation since 2013. Although data on research projects had been processed before, their quality and the scope of the monitored data were not sufficient.

Since 2013, we have continuously strived to increase the completeness and quality of the data. Our work during this period was based on access to the SQL database in R/W mode, where we performed qualitative adjustments to records using routines based on the update, delete (deduplication and removal of irrelevant records) and insert (batch data insertion) commands.

This software solution for data administration did not meet even the basic requirements for cybersecurity and was not resistant to accidental data loss or intentional manipulation. It was based on the caution and professional knowledge of the staff, which we have proven by the fact that we have not had any incidents in 10 years of operation.

In 2024, we switched to a new software solution that meets security requirements and already includes a basic administrator interface. However, since this interface is only used for minor corrections, we were forced to explore the possibilities of automating data quality management with the aim of:

- Ensuring the highest possible level of database completion, i.e. including data on research projects that are not included in the database in the form of batch data processing.
- Identifying incomplete and incorrectly filled records and ensuring their completion and correction.

METHODS. Our research began by summarizing knowledge and experience across European countries. Stempfhuber (2008) already states that quality is defined very differently in individual CRIS systems, but a large part of the efforts made on information quality are currently focused on standardization. The data model cannot cover the quality criteria, but their formalization is necessary. However, it must be said that the formalization of quality criteria has not yet occurred.

When we talk about quality within the CRIS system, we can emphasize several aspects:

- Information quality is understood as the correctness of the entered data. Here, authority files, dictionaries, thesauri, persistent identifiers, as well as tools for automatic control of the structure and values of the entered data are used.
- Data integration is increasingly based on the connection of related information systems and on the application of the principle of “once and for all”, when each data should be entered into the system only once.
- Information quality assurance is understood as a process including emphasized data personalization. Data is checked immediately after its creation based on set criteria, using existing data, but also using user feedback.

Azeroual & Schöpfel (Azeroual & Schöpfel, 2019) define data quality in terms of suitability for use and state that it is not an absolute value, but a dynamic concept. The result is user satisfaction and acceptance of the system by them.

The authors further state that even small errors can make data unusable. The following specific quality problems frequently occurring:

- Missing values,
- Incorrect data caused by user error,

² <https://www.skcris.sk>

³ <https://eurocris.org/services/main-features-cerif/>

- Duplicate records,
- Inconsistent data formats,
- Logical contradictions between data.

In order to identify and eliminate quality problems, new techniques and methods of data cleaning and monitoring are needed.

The literature mentions a number of approaches and technologies suitable for this purpose.

One option is the use of the Data Profiling technique (Azeroual et al., 2018). This is an automated data analysis process, the aim of which is to: understand the structure of data, detect anomalies, identify inconsistencies, find relationships between attributes and support data cleaning.

The architecture of a specific software solution based on a combination of data lake and data wrangling is offered by Azeroual, O. et al. (Azeroual et al. 2022). Gurney, T. (Gurney, 2022) uses the example of the PURE system to connect the data-quality framework with specific CRIS functionality. Lappalainen et al. (Lappalainen et al., 2023) describe the use of data-wrangling and machine-learning techniques. Specific solutions are also offered in the article by Peroni, S.; Rizzetto, E. (Peroni & Rizzetto, 2025).

Our approach took into account the specifics of data collection on science and research in Slovakia, established procedures and required data collections, as well as the capacity capabilities of our five-member team. We set ourselves the following tasks:

1. Determine which objects will have priority when setting up quality management. We made quick decisions on this task, also based on previous experience with the operation of the CRIS system.

The choice fell on the register of research projects. There were several reasons: First of all, the SKCRIS system is mainly single-level. Only projects from the systems of Slovak grant agencies via API interfaces are validated. However, not all agencies have their own systems. Validation by the entity entering data, if it is a research organization, is not implemented. For this reason, we must address not only the correctness and completeness of the data entered. It is also necessary to create a method for entering data on research projects that are not yet included in the system, but are relevant and should be in the system. It concerns international cooperation projects, but also some domestic research activities.

Another challenge is the researcher register. Researchers are currently registering in this register themselves, or they are registered by an employer organization. There is not a single API interface with another system in production yet, although we have several in Slovakia. Within the researcher register, we also enter missing links of individual researchers to projects (i.e. *Person V is the researcher of Project P*). These links are created automatically only in the case of successful import of grant agency projects via the API. Their successful automatic creation is mainly limited by the frequent absence of unique identifiers.

2. Identify the main areas of errors and problems that we will address. We have set the following most important tasks:

- Batch import of data on research projects
- Assignment of projects to researchers by batch way
- Checking incomplete and incorrect data and resolving duplicates

3. Prepare and deploy software solutions designed for data quality management. We partly solved this task with the software solution supplier SK CRIS. A separate sub-task, for which we did not need the cooperation of the supplier, was the automation of data preparation for SK CRIS for batch import, where we also addressed the possibility of using artificial intelligence.

RESULTS AND DISCUSSION. Let's take a closer look at the register of research projects and the data that enters it. Grant agencies are obliged by Act No. 172/2005 Coll. to enter data on completed projects. The SK CRIS information system provides grant agencies with an API integration interface that allows them to fulfill this obligation. The alternative is to provide data in another way (usually MS Excel), but in this case the task of importing the provided data is transferred to the system operator.

A separate area is international cooperation in science and research. A large number of grant schemes from various entities intended for research organizations is identified across Europe.

We obtain data on international cooperation from public sources (websites of funding providers) or from research organizations. The focus of processing data on international cooperation lies with the system operator. The categorization of international cooperation projects is governed by a code list that contains selected grant schemas (for example, COST, ERASMUS +, HORIZONT 2020, HORIZONT Europe and others). A significant category is Other international cooperation, intended for projects that we cannot classify by funding source.

Data on completed projects are also provided separately by universities, which have a strong motivation to provide complete and high-quality data, since research activity and obtained grants are one of the funding criteria.

Research projects by data source and method of entry into the system are shown in Fig. 1.

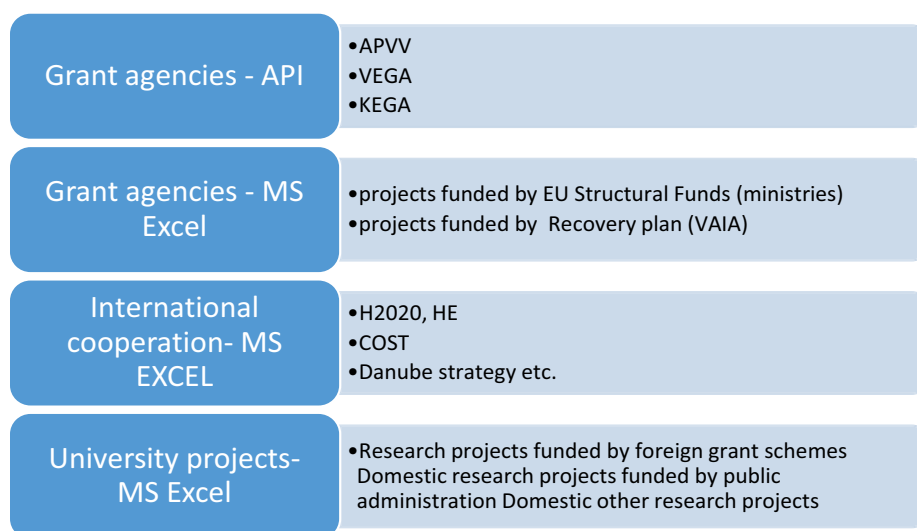


Fig. 1. SK CRIS projects by funding source

The method of obtaining projects shown in Fig. 1 requires a lot of work on the administrator's part, which we had to automate at least partially. In addition to MS Excel, we used the free POSTMAN⁴ API platform in the basic version, which allows work on system data without the need to access the SQL database environment directly. In cooperation with the supplier SK CRIS, a template was created for batch import of project data directly from the xlsx format.

Another problem that we solve when registering data on research projects is the research team. In practice, this involves creating links between a research project and a researcher with the assignment of semantics (most often one of the following roles is assigned: the responsible researcher or researcher). Ideally, if all involved stakeholders use consistently the unique researcher identifier (ORCID, SCOPUS_Id, etc.), the

⁴ <https://www.postman.com>

process would run smoothly. Since in Slovakia the identifiers not used universally, the identification based on a combination of fields. For example, first name+last name+date of birth or first name+last name+email. In the case of matching based on field combinations, the matching is not always successful. In such a case, it is necessary to find the correct researcher based on the error message from the system.

During imports, we process the entire list of unmatched researchers and use POSTMAN again to supplement the data and links.

The result is records of research projects that can be used for various purposes (Fig. 2). Data on research teams is used, for example, to prepare statistics in the field of gender equality.

As already mentioned (Fig. 1), we obtain a lot of data in MS Excel and it is our task as the operator to import it into SK CRIS. The batch of data on research projects for SK CRIS, which we received, about 1000 records, was characterized by a generally diverse quality of data. The data on the field of science and technology, which is a mandatory field of the SK CRIS data model, was completely missing from the records.

For specification of the scientific disciplines in the SK CRIS system, the three-level code list of R&D fields⁵ is used. The code list is based on the Frascati manual (OECD)⁶ and is defined by valid Slovak legislation.

Project code
APVV-23-0265

Acronym
3Driver

Project duration
01.09.2024 - 30.06.2028

Keywords
point cloud, river system, hydromorphological quality, habitat, deep learning, classification

Call name
25.09.2023 - APVV Verejná výzva na predkladanie žiadostí na riešenie projektov výskumu a vývoja v jednotlivých skupinách odborov vedy a techniky – VV 2023

Allocated amount
169514.00 EUR

Department of Science and Technology
Natural Sciences, Mathematical Sciences, Informatics and Cybernetic Sciences / Earth sciences and environmental sciences (incl. earth resources) / Physical geography and geocology

Character of R&D
basic research

Program type, Financial source
SR&DA (APVV) Public Call

3D technology and machine learning methods for remote environmental monitoring and assessment of riv...

Expand full heading

SKCRIS > Register of projects > Project detail

Riparian zones are among the most biologically diverse and productive ecosystems on Earth. They are shaped by underlying physical processes associated with river flow, including erosion and deposition of sediment, periodic inundation, and groundwater-surface water exchange. High-resolution point cloud topography data offers new opportunities for de...

View full abstract

Researchers	
Name ↓	Role
Hamid Afzali	PhD Student
Katarína Bónová	Investigator
Michal Gallay	Investigator
Jaroslav Hofierka	Investigator
Ján Kaňuk	Investigator
Ashraf Md	Researcher
Lukáš Michaleje	Principal investigator
Soheyl Moradi	Investigator
Zuzana Pazúrová	Investigator
Miloš Rusnák	Principal investigator

1-10 of 13

10 ▼

Fig. 2. Record of Research project in the SK CRIS Project Register

⁵ Číselník Odborov Vedy a Techniky. <https://www.vedatechnika.sk/sk/veda-sr/odbory-vedy-techniky/> R&D – Research and Development

⁶ Frascati Manual. Guidelines for Collecting and Reporting Data on Research and Experimental Development. 2015. <https://www.oecd.org/sti/inno/Frascati-Manual.htm>.

An example of three levels from the Slovak/SK CRIS code list of R&D fields is provided by way of illustration:

Technical Sciences → Industrial biotechnology → Biomaterials

In practice, the R&D field is defined by the entity that enters the data. As a rule, this is the project organization, but certain corrections can also be made by the grant agency. If this data is missing, it is necessary to supplement it within the framework of data quality management, otherwise it is not possible to import (incomplete) data. We therefore addressed how to achieve the required data quality and complexity as easily as possible.

The task was therefore whether it is possible to use AI tools to determine the probably correct science and technology field based on the project name and its annotation (Zendulková & Azeroual, 2026) and based on the entered code list.

The result using GEMINI was positive and is shown in Fig. 3.

	A	B	C	D
1	Názov projektu	Úroveň 1	Úroveň 2	Úroveň 3
2	Sociálne stiahnutie v bežnej a klinickej populácii: porozumenie komplexným vzťahom medzi sociálnymi, psychologickými a neurobiologickými procesmi	050000 SPOLOČENSKÉ VEDY	050100 psychologické vedy	050102 klinická psychológia
3	Pushing Boundaries in Molecular Materials by Combining Various Functionalities	020000 TECHNICKÉ VEDY	020500 materiálové inžinierstvo	020503 kompozity
4	Genetic profile variability in urinary bladder carcinoma; implications for diagnostics and therapy.	030000 LEKÁRSKE VEDY A ZDRAVOTNÍCKE VEDY	030100 základné lekárske vedy (vrátane z	030107 lekárska onkológia
5				

Fig. 3. Categories of R&D fields according to the index, assigned by artificial intelligence

In the framework of data quality management, it is essential to determine priorities, *select objects* that will be checked as a priority. We focused primarily on the basic objects of the SK CRIS, which are research projects, researchers, R&D organizations and research results.

We did not mention organizations at all in the article. We do not address the quality of records on organizations due to the connection of the SK CRIS Organizations Register with the national reference Register of Legal Entities. We therefore consider the data as correct.

As for research results, SK CRIS is not primarily intended for their general registration. The Central Register of Publication Activities of Universities⁷ serves this purpose. SK CRIS is connected to this register via API interfaces and automatically creates links between publications and researchers and organizations, and research projects only to the extent that this data is filled in.

However, research results are entered into SK CRIS by organizations that are not universities. The result is hundreds of thousands of records, the control of which is a challenge for the future.

So far, the subsequent check after data entry into the database has been mentioned. When entering data into SK CRIS via a form and API, *data checking* is built in: mandatory fields are specified, templates for items with a fixed structure, and values selection from controlled lists and code lists.

During manual subsequent check, we use MS Excel or SQL commands to obtain data directly from the database. The data output is subject to filtering, where we identify unfilled or insufficiently filled fields, for example, a dash or one character in the case of a mandatory field.

We have partially solved the solution of duplicates within the forms, but the check cannot catch situations where the user intentionally enters data repeatedly, since he is not sure of the success of the entry process. It is also difficult to identify duplication in the case of a differently entered project name. A typical example

⁷ <https://www.crepc.sk>

of a difficult-to-identify duplication is a situation where the project name in one record is in Slovak and in the second record is in English.

We consider data quality management to be a continuous process, the result of which is a greater number of quality records in the database. This is also evidenced by the overview of the number of records on research projects entering SK CRIS by individual years, which can be found in Table 1.

Table 1. Overview of the number of research projects by year

Year	Number of projects started in the monitored year	Number of projects in solution in the monitored year
2016	1400	3851
2017	1279	4135
2018	1290	4313
2019	1362	4652
2020	1434	4683
2021	1390	4794
2022	1496	4971
2023	1358	4898
2024	2015	5404
2025	1574	5796

CONCLUSION. Data quality management in a CRIS is a continuous process that must reflect both the characteristics of the data sources and the operational capacities of the system administrator. The experience gained from the operation of SK CRIS demonstrates that data quality cannot be ensured solely through validation at the point of data entry. In an environment where research information is collected from multiple heterogeneous sources, additional mechanisms for data validation, cleaning, enrichment, deduplication and monitoring are required.

The implementation of the new SK CRIS solution has enabled a gradual transition from direct database interventions towards safer and more systematic data management. The tools and procedures described in this paper, including API-based processing, batch import, automated identification of incomplete records and the use of controlled vocabularies, represent important steps towards reducing manual intervention and improving the completeness and consistency of research information.

Particular attention was paid to the Register of Research Projects, where data quality is strongly affected by heterogeneous data sources, incomplete use of persistent identifiers and missing mandatory attributes. The experience with linking researchers to projects confirms the importance of persistent identifiers such as ORCID for reliable entity matching. At the same time, the experiment with AI-assisted classification of research projects indicates that artificial intelligence can support data enrichment in cases where structured information is missing. Such automatically generated values, however, should be considered recommendations and remain subject to appropriate validation before their inclusion in the authoritative CRIS data.

The results also show that no single software tool can address all dimensions of CRIS data quality. An effective approach requires a combination of preventive validation during data entry, automated processing and quality checks, controlled vocabularies and identifiers, as well as subsequent monitoring and expert supervision. The selection of tools should also take into account the size and capacities of the team responsible for system operation.

Future work will therefore focus on extending automated quality-control mechanisms to other SK CRIS entities, particularly researchers and research outputs, improving duplicate detection and entity matching,

and further exploring AI-supported methods for identifying, correcting and enriching problematic records. A further objective is to reduce the amount of manual data processing while maintaining human supervision in cases where automated decisions cannot provide sufficient reliability.

The SK CRIS experience confirms that improving data quality is a long-term combination of technological, organisational and methodological measures. Developing integrated software support for these processes is therefore essential for maintaining a reliable national research information infrastructure and for increasing the usability of CRIS data for reporting, evaluation, for analysis and research policy.

FUNDING. This paper was prepared with the support of project No. 401101DVH7, Support for Slovakia's Participation in the European Research Area II (SK4ERA II) cofounded by EU Programme Slovakia.

REFERENCES

- Azeroual, O., & Schöpfel, J. (2019). Quality Issues of CRIS Data: An Exploratory Investigation with Universities from Twelve Countries. *Publications*, 7(1), 14. <https://doi.org/10.3390/publications7010014>
- Azeroual, O., Saake, G., & Schallehn, E. (2018). Analyzing data quality issues in research information systems via data profiling. *International Journal of Information Management*, 41, 50–56. <https://doi.org/10.1016/j.ijinfomgt.2018.02.007>
- Azeroual, O., Schöpfel, J., Ivanovic, D., & Nikiforova, A. (2022). Combining Data Lake and Data Wrangling for Ensuring Data Quality in CRIS. *Procedia Computer Science*, 211, 3–16. <https://doi.org/10.1016/j.procs.2022.10.171>
- Gurney, T. (2022). The changing scope of data quality and fit for purpose: evolution and adaption of a CRIS solution. *Procedia Computer Science*, 211, 201–205. <https://doi.org/10.1016/j.procs.2022.10.192>
- Lappalainen, Y., Lassila, M., Heikkilä, T., Nieminen, J., & Lehtilä, T. (2023). Migrating 120,000 Legacy Publications from Several Systems into a Current Research Information System Using Advanced Data Wrangling Techniques. *Publications*, 11(4), 49. <https://doi.org/10.3390/publications11040049>
- Peroni, S., & Rizzetto, E. (2025). A tool for validating and monitoring bibliographic data in open research information systems: The OpenCitations collections. *CEUR Workshop Proceedings: Proceedings of the 21st Conference on Information and Research science Connecting to Digital and Library science*, 1–13. <https://hdl.handle.net/11585/1050675>
- Stempfhuber, M. (2008). Information quality in the context of CRIS and CERIF. <https://dspacecris.eurocris.org/handle/11366/331>
- Zendulková, D., & Azeroual, O. (2026) Practical Prompt Engineering for AI-Supported Metadata Processing in CRIS. Conference paper. 17th International Conference on Current Research Information Systems (CRIS2026). <https://dspacecris.eurocris.org/handle/11366/10292>
- Zendulková, D., Gavurníková, G., & Turňa, J. (2026). The Ecosystem of State Support for Science in Slovakia and Scientific Data. In O. Azeroual (Ed.), *Technical Foundations and Applications of Trustworthy AI Systems* (pp. 225-262). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-2600-1433-2.ch008>

Зендулкова Даница,

Словацький центр науково-технічної інформації, Братислава, Словаччина
ORCID: 0000-0003-2487-0177, e-mail: Danica.zendulkova@cvtisr.sk

ПРОГРАМНІ ЗАСОБИ ДЛЯ УПРАВЛІННЯ ЯКІСТЮ ДАНИХ У СИСТЕМІ CRIS

АНОТАЦІЯ. Словацька інформаційна система поточних досліджень (SK CRIS) інтегрує дослідницьку інформацію з кількох різномірних джерел, що створює значні труднощі для забезпечення повноти, точнос-

ті та узгодженості даних. Узагальнено практичний досвід управління якістю даних, отриманий під час роботи SK CRIS, та представлено програмно-підтримувані підходи до покращення якості дослідницької інформації. Особлива увага приділяється дослідницьким проєктам, включаючи пакетний імпорт даних, зіставлення дослідника та проєкту, виявлення неповних та дублікатних записів, а також збагачення даних. Впроваджені рішення поєднують вбудовані механізми валідації, обробку на основі API за допомогою Postman, пакетну обробку даних XLSX та контрольовані словники. Також досліджується потенціал штучного інтелекту для призначення відсутніх дослідницьких полів на основі назв проєктів та анотацій. Результати показують, що ефективне управління якістю даних CRIS вимагає поєднання автоматизованих інструментів перевірки, обробки та збагачення даних з експертним наглядом та постійним моніторингом.

КЛЮЧОВІ СЛОВА: Сучасна дослідницька інформаційна система; Управління якістю даних; Очищення даних; Моніторинг даних; SK CRIS; Програмні інструменти.