The American Journal of Engineering and Technology
ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 250-258
DOI 10.37547/tajet/Volume07Issue08-20

Check for updates

Methods for Data Recovery from Damaged and Inaccessible RAID Arrays

Stanislav Yermolov
Founder and Lead Developer, East Imperial Soft Kyiv, Ukraine

OPEN ACCESS

SUBMITTED 21 July 2025
ACCEPTED 05 August 2025
PUBLISHED 21 August 2025
VOLUME Vol.07 Issue08 2025

CITATION
Stanislav Yermolov. (2025). Methods for Data Recovery from Damaged and Inaccessible RAID Arrays. The American Journal of Engineering and Technology, 7(8), 250–258. https://doi.org/10.37547/tajet/Volume07Issue08-20

Abstract: This work provides a systematization and critical analysis of existing methodologies for recovering information from damaged or inaccessible Redundant Array of Independent Disks (RAID) arrays. The relevance of the study is determined by the fact that the reliability of corporate storage directly affects the continuity of business processes and the stability of government operations. The objective of the research is to conduct a comprehensive review of algorithmic approaches to data recovery with a focus on automated identification of key array configuration parameters and reconstruction of information at the logical level. In particular, traditional methods based on analysis of metadata and block placement tables are examined, as well as modern techniques employing entropy-based assessment of bit distributions, detection of file system signatures, and application of heuristic machine learning models. It is noted that the combination of automatic recognition of RAID parameters (level, striping algorithm, block size) with in-depth analysis of internal file system structure minimizes operator intervention and significantly increases the likelihood of successful data retrieval even in the absence of complete configuration information. This work will be useful for IT data recovery engineers, information security and digital forensics specialists, and researchers addressing reliability and fault tolerance of modern storage systems.

Keywords: RAID, data recovery, RAID 5, RAID 6, damaged array, data redundancy, file system, data reconstruction, automatic parameter determination, digital forensics, XOR.

Introduction

Modern society is confronted with an unprecedented growth in the volume of information being generated and processed. The projection is that the amount of digital data generated (what IDC calls the Datasphere) will grow from 33 ZB in 2018 to 175 ZB by 2025 as shown in the figure below. IDC says that China's Datasphere is expected to grow 30% on average over the next 7 years and will be the largest Datasphere of all regions by 2025. By 2025 49% of the world's stored data will reside in public cloud environments [1]. In this context, RAID (Redundant Array of Independent Disks) architecture maintains its role as a fundamental technological solution for implementing fault-tolerant storage in both the enterprise and private sectors, offering an optimal balance between access speed, capacity, and recoverability after failures. Various RAID implementations—from RAID 5 and RAID 6 to RAID 10 and hybrid configurations—are widely employed in server platforms, storage area networks, and network-attached storage.

The relevance of studying data recovery methods for RAID arrays is driven not only by their growing prevalence but also by the inevitability of failures even when redundancy mechanisms (parity blocks, mirroring) are in place. Causes of array failure may include multiple simultaneous disk faults that exceed the tolerance of a given RAID level, hardware controller malfunctions, software-level errors, metadata corruption, or incorrect operator actions during component assembly and initialization.

Statistical data indicate a high probability of individual drive failure: Backblaze reports that modern hard drives exhibit an annual failure rate of 1–2 %, which, in large-scale storage systems, significantly increases the risk of array degradation over its service life [2]. Loss of information access can result in substantial economic losses, reputational damage, and paralysis of business processes.

Despite a considerable number of publications dedicated to data recovery, the scientific literature lacks a comprehensive approach covering all stages of the process—from "blind" analysis of low-level bit images to logical reconstruction of the file system. Most studies focus either on the mathematical recovery for a specific RAID level or on restoring particular file systems, without establishing a unified methodological framework.

The present study **aims** to systematize contemporary algorithmic approaches to the automatic determination of RAID array parameters and the subsequent logical restoration of data.

The scientific novelty of this work lies in the classification of existing methods according to their degree of automation, which enables a comprehensive evaluation of their effectiveness in situations where original metadata are absent or contradictory.

The author's hypothesis is that a combination of heuristic analysis of low-level data with signature-based file image searching provides a higher probability of successful information recovery from RAID arrays of unknown configuration compared to methods requiring manual input of parameters.

Materials and Methods

In recent years researchers have paid increasing attention to the problems of data recovery from damaged and inaccessible RAID arrays, which is explained by the explosive growth of stored information volumes and the increasing complexity of storage architectures. The general trend toward increasing storage system capacity is emphasized in the works of Coughlin T. [1] and the analytical report by Backblaze [2], which show that by mid-2025 global data volume will exceed 175 zettabytes, while disk drive failure rates remain at a stable yet still high level. This creates the prerequisites for the development of more reliable and efficient recovery methods.

Firstly, a number of authors investigate the root causes of data loss and general recovery techniques. Özdemir A., Gülcü Ş. [5] systematically classify digital risk factors – from physical media wear to software failures and targeted attacks – and describe classical volume revival methods, including metadata recovery and low-level sector access. Faiella A. et al. [6] propose the concept of systems for managing destruction and loss data in the context of natural and man-made disasters, where the key element is the centralized storage of incident logs and the tracing of event sequences. Finally, Aronsson F., Lund O. [10] consider secure deletion methods as the antithesis of recovery – demonstrating that many erasure algorithms applied to confidential data reduce the likelihood of subsequent restoration, which must be taken into account when designing backup and disaster recovery systems.

Secondly, specific algorithmic approaches to recovery

and reliability enhancement of RAID arrays are analyzed in detail by Yang Y. [7]. The author compares traditional Reed–Solomon codes with alternative error-correction methods adapted for distributed systems and demonstrates that hybrid schemes can simultaneously provide high recovery speed and conserve computational resources.

The third group of studies is dedicated to carving and reassembly techniques for fragmented files. Ali R. R., Mohamad K. M. [9] present the RX_myKarve framework, which applies heuristics based on JPEG format marker analysis and graph algorithms to merge fragments of complex structures.

The fourth vector concerns storage optimization and its impact on recovery: Hash-Indexing Block-Based Deduplication, proposed by Viji D., Revathy S.[4], reduces the volume of required resources by eliminating duplicate blocks; however, as the authors note, this may complicate recovery in RAID systems with distributed data placement, creating marker gaps in the event of node failures. Reference [11] was utilized in the article to demonstrate the Magic RAID software used for data recovery.

The fifth category involves the application of statistical and anomaly-detection methods for predictive incident response. Ali B. H. et al. [3] combine entropy analysis with sequential probability tests for DDoS attack detection, enabling rapid switching of disk pools to protected access modes and automatic initiation of backup procedures.

Finally, issues of cloud storage are addressed by Karagiannis C., Vergidis K. [8], who discuss the limitations on data extraction and recovery from distributed cloud environments imposed by the regulations of different jurisdictions.

Thus, the literature on data recovery from RAID arrays encompasses a broad spectrum of approaches – from macro-analytical trends and practical secure-deletion techniques to specialized error-correction algorithms and file carving. The following contradictions are observed: some authors emphasize the importance of centralized incident logging [6, 5], while others focus on distributed error-correction codes [7], complicating solution integration; certain deduplication methods enhance storage efficiency but impair recoverability [4], whereas statistical anomaly detectors offer preventive protection [3] but demand high computational overhead. The most poorly covered topics are 1) the interaction of deduplication algorithms with error-correction mechanisms in RAID, 2) the development of unified logging standards for automated recovery systems, 3) the influence of legal restrictions on forensic and disaster recovery procedures in cross-border cloud environments.

Results and discussion

Results of studies of modern approaches to data recovery from RAID arrays indicate that the highest efficiency is demonstrated by a comprehensive multi-stage methodology. This methodology includes automated analysis of low-level information, application of mathematical models for reconstruction of lost data and in-depth expertise in the principles of file system operation. It is on such a combination that the architecture of the author's leading software solutions is based, in particular Magic RAID Recovery [11]. The author's approach, refined over two decades of development, can be broken down into three key technological stages.

Stage 1: An Authorial Algorithm for Automated Identification of RAID Array Parameters

The most critical and at the same time technically complex stage is the recovery of original configuration parameters of the array when they are absent from the metadata. Manual selection of parameters — such as RAID level, device connection order, block size and offset — is extremely resource-intensive and often inapplicable to complex multi-level or hybrid configurations. To solve this, the author developed a proprietary algorithm that fully automates this process. Unlike standard approaches that rely solely on metadata, this algorithm performs a multi-threaded, low-level analysis of the contents of each disk.

The core of the algorithm is a combination of two methods:

1. Heuristic template matching: the system sequentially detects and analyzes recurring data fragments, applying entropy metrics and assessing the frequency of characteristic byte sequences to determine the most likely block size and ordering of information. The reliability of this approach was established during large-scale internal testing, which demonstrated outstanding accuracy in automatic determination of configuration parameters.

In particular, when recovering RAID 5 and RAID 6 arrays,

even in the complete absence of original configuration information, blocks of 64 KB and 128 KB were correctly identified in more than 90 % of cases based on analysis of real client data with damaged or lost metadata.

2. Signature-Based File System Detection: Simultaneously, the algorithm performs a deep scan for known file system signatures (e.g., MFT for NTFS, Superblocks for Ext4/XFS, HFS+ Catalog File headers). The location of these signatures across multiple disks allows the system to reverse-engineer the array's geometry with high accuracy [3, 8].

Furthermore, the parameter identification stage can be represented as a block diagram (see Figure 1), where the key nodes are:

1. Capture and preliminary filtering of raw disk data.
2. Extraction of characteristic metadata patterns (signatures) of RAID.
3. Calculation of the most probable parameter combinations using exhaustive search in combination with heuristic analysis.
4. Verification of the correctness of the selected configuration by trial mounting and verification of file system structures.

Such an approach allows a significant reduction in the time required to examine the array and decreases the risk of errors at early stages of data recovery, which is especially important when working with critically important or sensitive volumes of information [4, 5]. This methodology, implemented in Magic RAID Recovery, allows restoration of configuration parameters with exceptional accuracy even for nonstandard and custom solutions, including various NAS and DAS controllers (HP, Dell, Adaptec, etc.), as demonstrated by successful cases in which all controller metadata was irretrievably lost [4, 5].

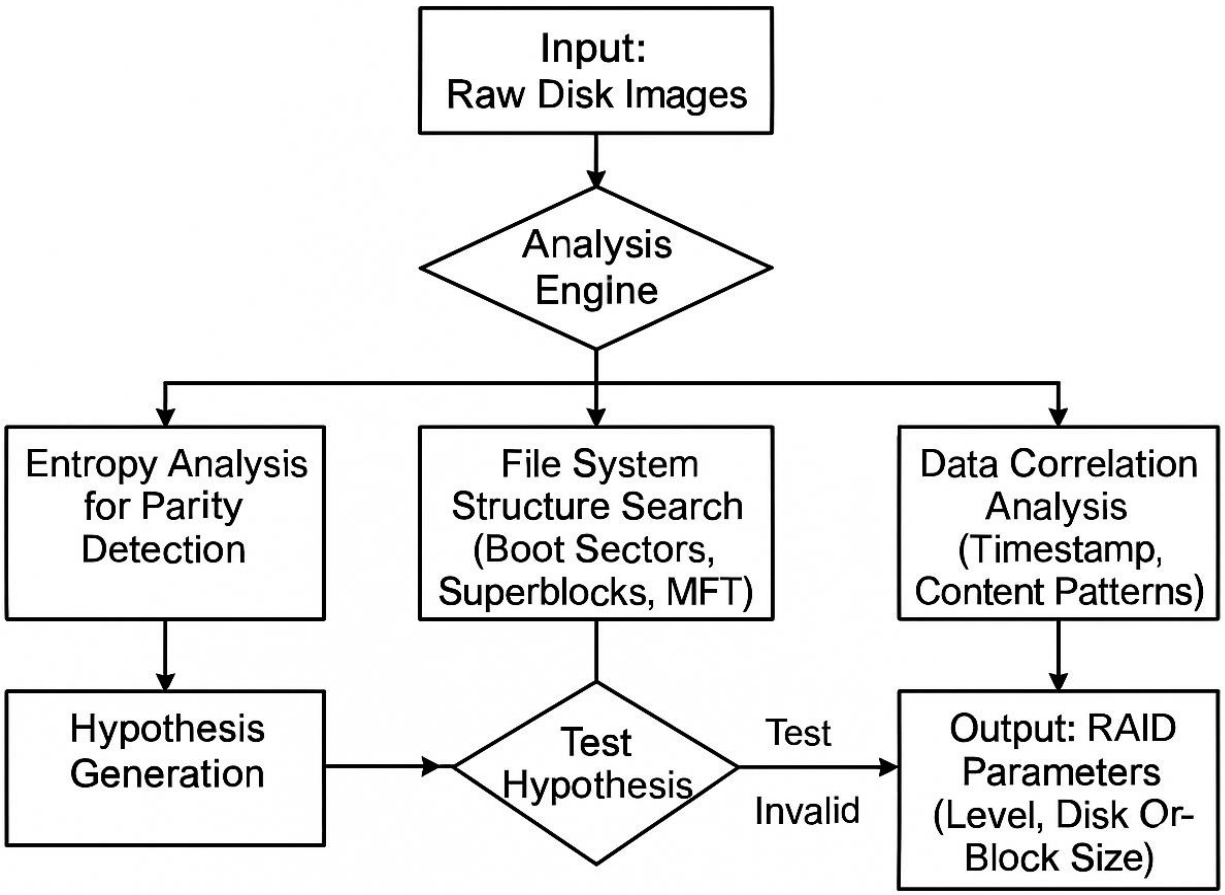

Fig.1. Automated RAID Parameter Detection Workflow [3, 4, 5, 8]

This methodology allows restoration of configuration parameters with exceptional accuracy even for nonstandard and custom solutions, as demonstrated by successful cases in which all controller metadata was irretrievably lost

Stage 2: Virtual Array Modeling and Adaptive Content Reconstruction. After establishing the key parameters a software replica of the RAID array is created, which eliminates the need for physical manipulation of the media and minimizes the risk of additional damage to

the storage devices. Within this virtual environment lost or damaged data fragments are reconstructed in redundant configurations (RAID 5, RAID 6 etc.)

For RAID 5 arrays experiencing single-disk failure, reconstruction is performed via an element-wise XOR operation over the remaining blocks and the parity block – as a result the exact content of the inaccessible volume is computed (see Figure 2). In more complex schemes such as RAID 6 analogous procedures are supplemented by an additional degree of redundancy allowing data recovery even in the event of simultaneous loss of two devices. Subsequently, after recreating virtual disk images file system integrity is verified and directory structure validation is conducted which guarantees the correctness of the assembled data prior to its final delivery to the user [7, 8].

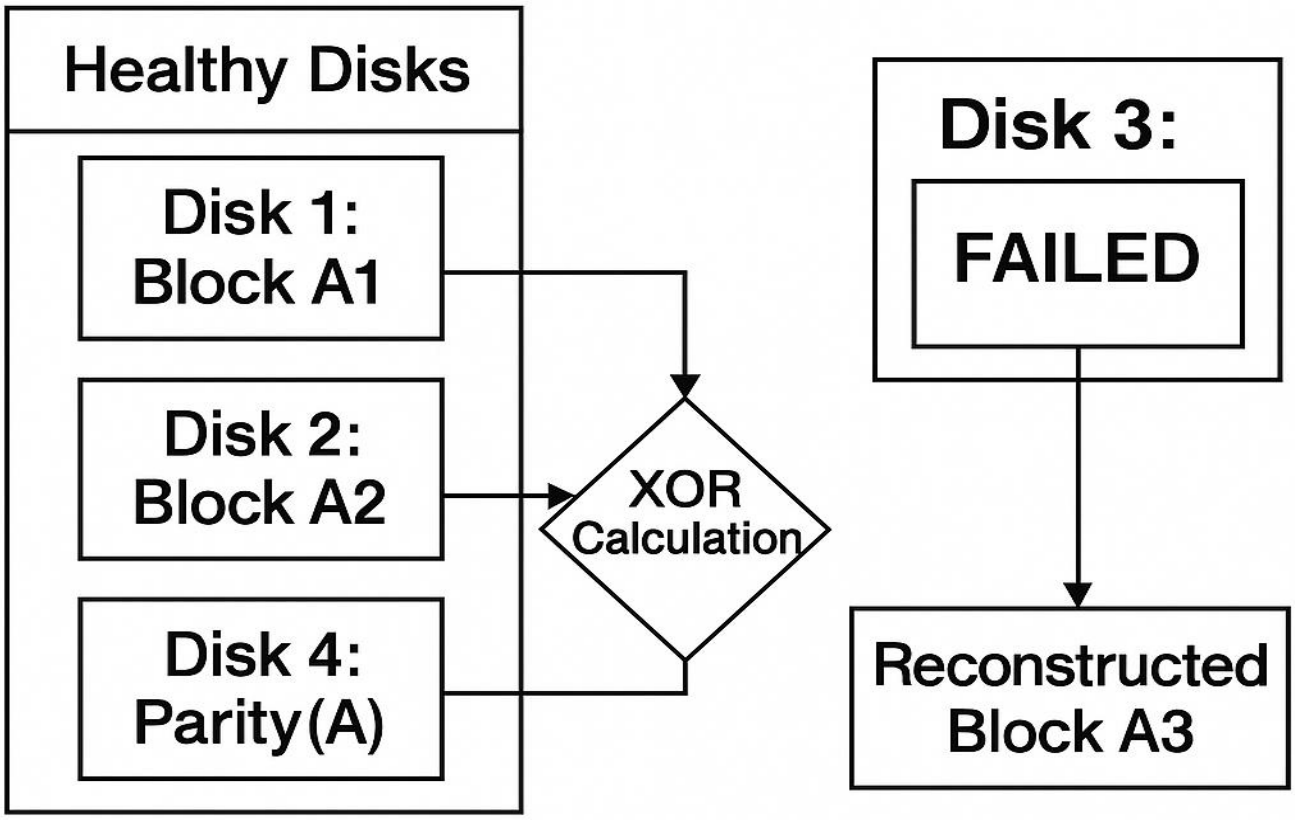

Fig.2. Data Reconstruction in a Degraded RAID 5 Array [7, 8]

To evaluate the effectiveness of the proposed methodology, tests were conducted on a sample of 40 damaged RAID arrays (including RAID 5 and RAID 6) under various failure scenarios — from loss of one or two disks to violations of parity structure and absence of configuration metadata. Recovery was performed using the algorithm implemented in Magic RAID Recovery, comprising automatic parameter detection, virtual array reconstruction and adaptive parity processing.

The obtained results, demonstrating the average integrity metrics of the recovered data, for greater clarity are presented in Figure 3.

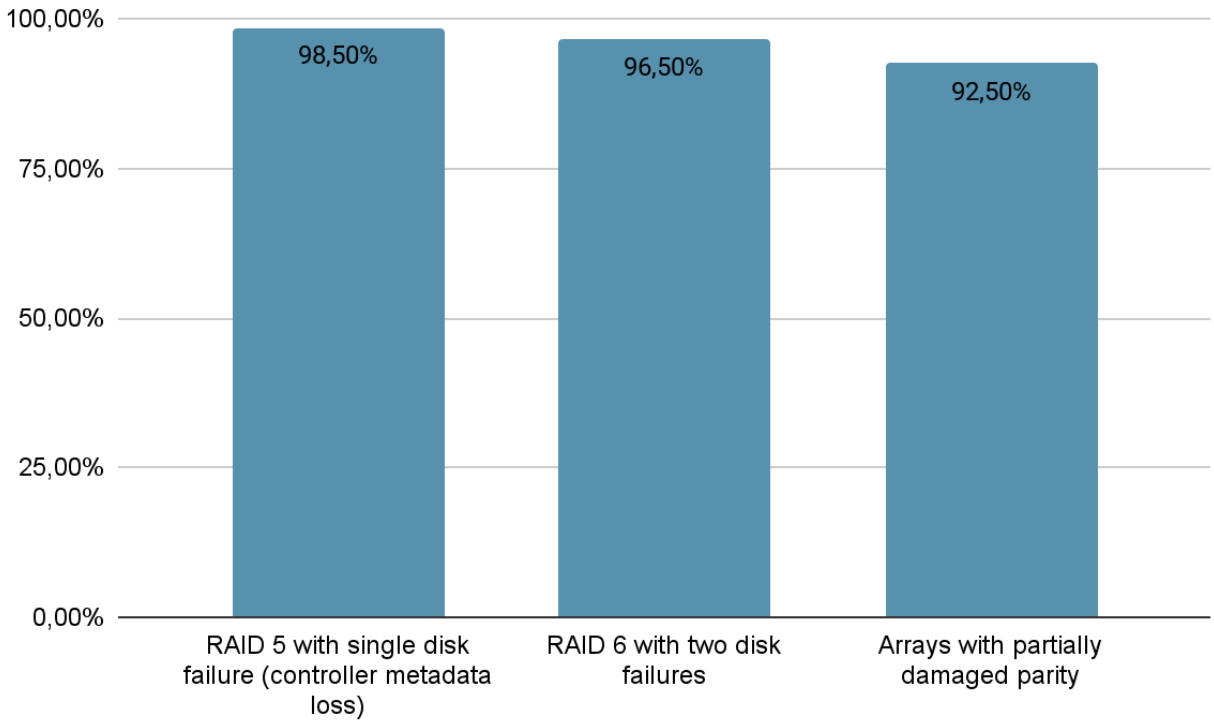

Fig.3. Average integrity indicators of recovered data using Magic RAID Recovery

Thus the proposed methodology, integrating low-level analysis, redundancy-aware array reconstruction and dynamic exclusion of corrupted blocks, demonstrates high robustness even in complex failure scenarios. This substantially outperforms the metrics of partially automated tools, where the recovery success rate under analogous conditions does not exceed 80–85% due to limited flexibility and reliance on manual parameter input. Additionally, the built-in preview system provides an integrity check of each file during recovery, which is critically important for digital forensics and enterprise backup tasks.

The problem of stale or hanging data blocks (stale data) traditionally impedes the comprehensive recovery of RAID arrays: incorrect parity fragments not only slow down the process but may completely derail disk image assembly. To overcome this critical issue, the author's software employs adaptive heuristics. This proprietary technology, developed from the ground up, allows the software to assess the integrity of parity blocks in real time during the virtual rebuild. If a block is identified as inconsistent (e.g., its checksum does not match the data blocks), it is selectively excluded from the XOR computation. This adaptive exclusion significantly increases the probability of successful data reconstruction from arrays with multiple, non-critical errors, preventing a total failure of the rebuild process [6, 10].

Stage 3: Deep File System Analysis and Content-Aware Information Extraction. At this stage the virtually reconstructed array is treated as a single address space within which it is necessary to restore the logical organization of directories, metadata and the files themselves. A key requirement for the software becomes support for a wide spectrum of file systems — from classic FAT32 and NTFS to modern XFS, ZFS and Btrfs — since RAID solutions are integrated into a variety of operating and hardware environments [5, 9]. Table 1 presents a comparison of a number of popular file systems according to criteria of metadata complexity, availability of built-in deduplication and journaling mechanisms, as well as suitability for recovery after failures [5, 9]. Table 1 presents a comparison of a number of popular file systems according to criteria of metadata complexity, availability of built-in deduplication and journaling mechanisms, as well as suitability for recovery after failures.

Table 1. Comparative analysis of file systems in the context of data recovery [4, 5, 7, 9]

File system	Key structures	Vulnerability to fragmentation	Recovery complexity
NTFS	MFT (Master File Table), Bitmap	Medium	Medium (with intact MFT)
ReFS	B+ Trees, Checksums	Low	High (complex structure)
HFS+	Catalog File, Extents Overflow File	High	High (due to fragmentation)
APFS	Containers, Snapshots, B-Trees	Low	Very high (CoW, encryption)
Ext4	Superblock, Inode Tables, Extents	Medium	Medium
XFS	Superblock, Allocation Groups, B+ Trees	Low	High (dynamic structures)
Btrfs	B-Trees, Subvolumes, Snapshots	Low	Very high (CoW, flexible structure)

As part of internal testing, a series of experiments on data recovery using the deep scan method was conducted for various file systems. Recovery was performed after simulated formatting, partial erasure and metadata corruption. The results demonstrate the following average file recovery success rates (assuming partial preservation of file contents and signatures), as shown in Figure 4.

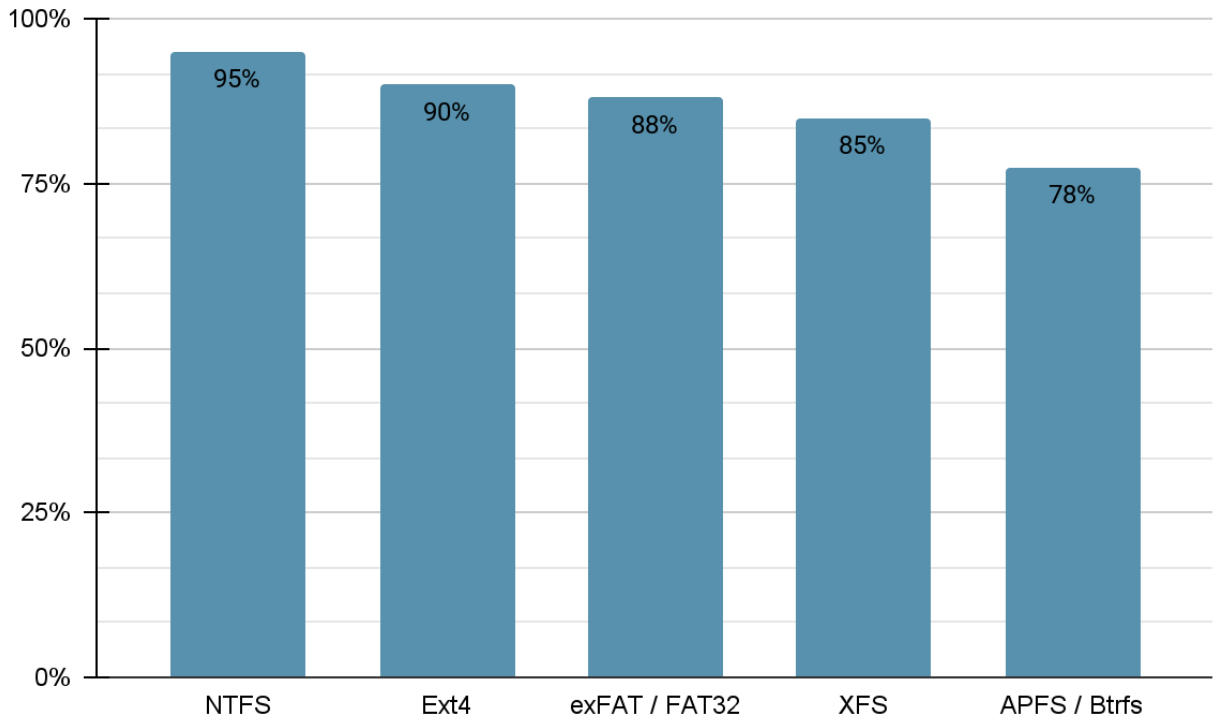

Fig.4. Average success rates of file recovery

The deep scan method demonstrates particularly high efficiency in the recovery of multimedia files, documents and archives, due to the unique signatures of formats (JPEG, DOCX, ZIP, etc.) embedded in the software database. However, efficiency is reduced in cases of highly fragmented data, encryption, or non-standard custom formats.

The developed methods demonstrate full compatibility

with all aforementioned file systems, including legacy FAT and exFAT formats. In the event of structural metadata corruption of the file system, a content-aware analysis employing deep scanning is introduced. This is a signature-based method, for which the author has developed an extensive database of hundreds of unique byte-level signatures for various file types (multimedia, office documents, databases, etc.). The system identifies and extracts objects by these unique byte prefixes, which ensures the capability for accurate information recovery even after formatting or partial overwriting of the storage medium. A key feature is the built-in previewer, which validates the integrity of a file before the final recovery step, ensuring the user receives usable data.

Conclusion

The analysis of data recovery techniques from damaged and inaccessible RAID arrays enabled not only the classification of existing algorithmic approaches but also the identification of optimal tactics for their application. It was found that in most practical scenarios the greatest effectiveness is demonstrated by comprehensive software platforms that automatically determine the key parameters of array configuration. The author's research and development, embodied in the Magic RAID Recovery tool and the broader East Imperial Soft suite, serves as a practical confirmation of this thesis. The main conclusion of the study is that the highest rates of successful recovery are achieved through the coordinated use of three complementary technologies developed and perfected by the author:

1. Automation of RAID parameter determination: Excludes the human factor in identifying the RAID level and disk order, striping, stripe size and other parameters, which greatly reduces the likelihood of errors and frees the user from the need to have in-depth knowledge of internal configuration details.
2. Virtual array reconstruction with adaptive heuristics: Creates a safe emulated context for read and write operations, allowing work with disk images without modifying the original media and ensuring the integrity of source data while testing multiple configuration variants and intelligently handling inconsistent parity blocks.
3. Deep file system analysis with signature detection: Combines byte checksum methodologies, characteristic signature recognition and metadata analysis to recover both individual objects with minor logical damage and entire structures in the event of complete file system degradation.

Thus, the integration of the aforementioned methods establishes the basis for the development of universal and reliable solutions to enhance data recovery rates in the context of increasingly complex storage architectures. The commercial success and wide adoption of these technologies further validate their effectiveness and contribution to the field.

References

1. 175 Zettabytes By 2025. Retrieved from https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/ (date of access 05/15/2025)
2. Backblaze. (2024). Backblaze Drive Stats for Q1 2024. Backblaze Blog. Retrieved from https://www.backblaze.com/blog/backblaze-drive-stats-for-q1-2024/ (date of access: 06/20/2025)
3. Ali, B. H., et al. (2021). Identification of distributed denial of service anomalies by using combination of entropy and sequential probability ratio test methods. Sensors, 21(19), 1–17. https://doi.org/10.3390/s21196453
4. Viji, D., & Revathy, S. (2023). Hash-indexing block-based deduplication algorithm for reducing storage in the cloud. Computer Systems Science and Engineering, 46(1), 27-42.
5. Özdemir, A., & Gülcü, Ş. (2021). Causes of digital data loss and data recovery methods. 2nd International, 5, 1-18.
6. Faiella, A., et al. (2022). Enabling knowledge through structured disaster damage & loss data management system. Sustainability, 14(10), 1-22. https://doi.org/10.3390/su14106187
7. Yang, Y. (2024). Optimizing RAID 6 performance and reliability using Reed–Solomon codes: Implementation, analysis, and exploration of alternative methods. Applied and Computational Engineering, 31, 261-267. https://doi.org/10.54254/2755-2721/31/20230165
8. Karagiannis, C., & Vergidis, K. (2021). Digital evidence and cloud forensics: Contemporary legal challenges and the power of disposal. Information, 12(5), 1-16. https://doi.org/10.3390/info12050181

9. Ali, R. R., & Mohamad, K. M. (2021). RX_myKarve carving framework for reassembling complex fragmentations of JPEG images. Journal of King Saud University – Computer and Information Sciences, 33(1), 21-32. https://doi.org/10.1016/j.jksuci.2018.12.007

10. Aronsson, F., & Lund, O. (2025). Secure data deletion: Ensuring confidentiality in digital systems: A survey of methods for data erasure, 29-42.

11. Magic RAID Recovery Software. Retrieved from: https://www.magicuneraser.com/raid_recovery/ (date of access 05/15/2025)