PROCEEDINGS OF SCIENCE

IHEP Docs: A Metadata-Based Unstructured Document Management System

QI Mengyao[a], WANG Li[a,b,c], LUO Qi[a], SUN Zhihui[a], QI Fazhi[a,b], and HOU Fengyao[a,b,*]

a Institute of High Energy Physics, Chinese Academy of Sciences (CAS),
No. 19B Yuquan Road, Shijingshan District, Beijing, China

b Department, University,
No. 1 Zhongziyuan Road, Dalang, Dongguan, China

c University of Chinese Academy of Sciences,
No.1 Yanqihu East Rd, Huairou District, Beijing, China

E-mail: qmy@ihep.ac.cn, wangli320@ihep.ac.cn, luoq@ihep.ac.cn, sunzh@ihep.ac.cn, qfz@ihep.ac.cn, houfy@ihep.ac.cn

The rapid growth of the volume of unstructured data brings great challenges to data storage, data security, data management and data utilization. The effective management of unstructured data has become a strategic need in the process of digital transformation of each organization. To meet with the urgent management need for massive document data scattered application systems and personal devices in the process of research, engineering construction and management, we developed and deployed a document management system (IHEP Docs) in institute of high energy physics of the Chinese Academy of Sciences (IHEP, CAS). This paper introduces its construction background, technical architecture, functional characteristics and the situation of application and operation in IHEP. At last, the future deployment plans for IHEP Docs are proposed to meet with the new and further requirements of document management and documents collaboration in IHEP.

International Symposium on Grids & Clouds (ISGC) 2023 in conjunction with HEPiX Spring 2023 Workshop, ISGC&HEPiX2023
19 - 31 March 2023
Academia Sinica Taipei, Taiwan

*Speaker

PoS(ISGC&HEPiX2023)017

1. Introduction

With the rapid development of information technology and its wide application, the storage and management of massive data has become a key issue, and it shows that unstructured data [1] accounts for about 95% of all data [2]. With characteristics of abundant data sources, large data volume, rapid data growth and difficult to manage in SQL (Structured Query Language) database, it brings great challenges to manage and utilize unstructured data effectively [3]. Masses of unstructured data, such as technical documents, papers, images, and videos, are deposited in personal computers or information application systems with lack of management and utilization. It has become an important application field to research and develop the management strategy of unstructured data and the related technologies [4]. In recent years, the development of big data, deep learning and natural language processing has made it possible to digitize unstructured data, thus build an intelligent and knowledge-based data platform that can be recognized, analyzed, and calculated by computers [5].

The Institute of High Energy Physics, Chinese Academy of Sciences (IHEP, CAS) is a comprehensive research base engaged in high energy physics, advanced accelerator physics and technology, advanced ray technology and its application, and has built a series of large-scale scientific facilities in China, such as Beijing Electron Positron Collider (BEPC) [6], China Spallation Neutron Source (CSNS) [7], High Energy Photon Source (HEPS) [8], etc.

The IHEP Document Management System（IHEP Docs）is designed and developed in order to solve the management and utilization problems of a large number of high-value unstructured data such as research documents, technical documents and management documents generated in the process of research activities, research management and construction of large-scale facilities, provide a secure, efficient and stable system of document management and document collaborative, and improve the ability of comprehensive management and office collaborative between departments and systems. At the same time, IHEP Docs is integrated with the information application system to realize the centralized storage and unified management of unstructured data in IHEP, empower the digital transformation, and support scientific research and the construction of large- scale scientific facilities in IHEP.

2. System Architecture

IHEP Docs is adopted the concept of unstructured data middle center and the unstructured content bus architecture which connects the two key platforms, which are the unified document management platform and the integrated application platform, and deployed with lightweight cloud services, in order to ensures the efficient connection of different services and the efficient unstructured data flow. The document management platform provides front-end services such as organization structure, document management, authentication & authority unified with IHEP SSO (single sign-on) integrated, etc. By the open standardized APIs of IHEP Docs, the application platform can provide an integrated application services such as WPS collaborative office [9], CAD online, OCR [10,11], collaborative sheet, customizable document workflows, and intelligent search engine, and anti-virus etc. It is expected to realize the centralized storage and unified management of the data of information systems and personal terminal devices and eliminate the data islands.

IHEP Docs provides users a unified workspace with one-stop document management, metadata services, document sharing & collaborations, content analysis, image recognition,

intelligent search, privacy management, log analysis and third-party application services. While realizing the users' personal document management, it provides a unified and comprehensive unstructured data service platform in IHEP. At the same time, deep learning, knowledge graph and other advanced technologies are adopted to realize the precipitation, mining, reuse, and re-creation from unstructured data, and provide users with knowledge services and useful tools, such as knowledge library, Wiki, Q&A, and knowledge community.

2.1Technical architecture

The hardware architecture of IHEP Docs adopts cluster including computing cluster, storage cluster and switch cluster. The system expect hardware is divided into four following layers, unified workspace, content bus service, Cloud native service and data management service (Figure. 1).

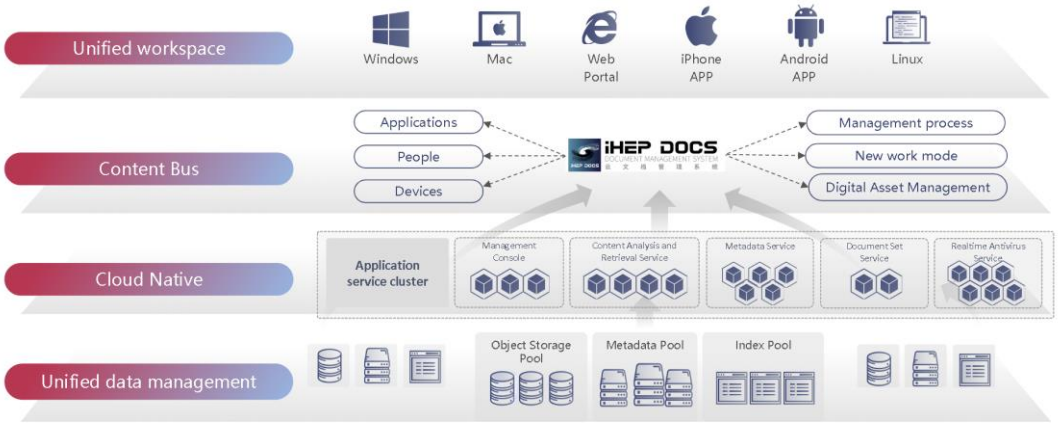

Figure 1 Technical architecture of IHEP Docs

2.1.1Unified workspace

IHEP Docs provides a unified workspace, which supports users to access the trans-region, trans-department data from different terminals anytime and anywhere, integrates applications such as IHEP SSO and third-party applications, such as WPS collaborative office, CAD online, and provide users with one-stop online collaborative office services such as document management, document collaboration, intelligent search, version management, collaborative sheet, and knowledge services.

2.1.2Content bus service

Based on content bus service (Figure.2), IHEP Docs can integrate the unstructured data of information system through the Open API interface, and realize a centralized storage, unified data management and effective document collaborations. The content bus service includes different functional modules, including fundamental services, data services, data applications, and data exchange.

The fundamental service is deployed on the high-performance hardware including computing cluster, storage cluster, network switches and security devices, adopts lightweight cloud computing technology to provide basic services such as virtualization, containers, databases, load balancing, security protection and middleware and support other services.

The data service includes search engine, document conversion, message services, log services, and content analysis services, and provides atomic-level capability support for IHEP Docs and open document services for other function modules.

The data application service provides rich application, including data aggregation, data storage, data management, document utilization, and maintenance, and supports different applications such as document upload, version management, intelligent search, document sharing, document circulation, log management, and system operation and maintenance.

The data exchange service provides standard interfaces, including standard APIs, single sign-on, application integration, data interaction and other modules. Integrated with the existing information systems and third-party applications, IHEP Docs has powerful scalability for the future development.

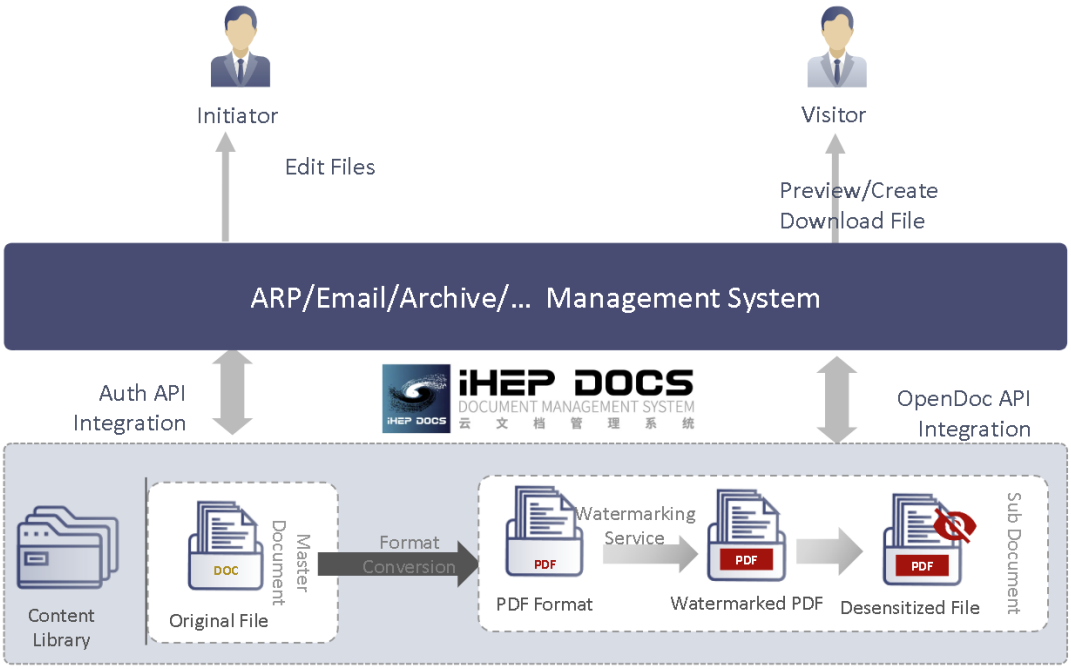

Figure 2 Open API and content bus service of IHEP Docs

2.1.3Cloud native service

IHEP Docs adopts lightweight cloud computing technology and runs in a self-built physical cluster environment. It is decomposed into the main modules and service modules with a separation of front-end applications and back-end data. All the modules can undertake independently functions, and be deployed, upgraded, and expanded independently to meet with the requirements of scalability, scalability, and maintenance.

2.1.4Data management

Data management service is divided into data storage, content processing, metadata generation, content service and so on according to the data flow process (Figure 3). To optimize the data management, a high-performance Ceph storage is used for data storage [12], and a high-performance content lake technology, which supports multiple homogeneous or heterogeneous storage clusters to bind to document libraries, is adopted to store and manage data to meet with

content classification and hierarchical storage. The content processing mainly involves the identification, extraction, and data retrieval of the key information of documents, picture files and video files, among which the process of data retrieval is the key to the performance optimization.

In order to promote a better performance in the process of data retrieval in massive unstructured data, a reverse index was adopted. The keywords were generated by data analysis in a certain order and then arranged into the data index, in which the order of index keywords and content coding is opposite to the one of data storage to avoid the performance bottleneck of retrieving one by one in order of document storage, and achieve a fast and comprehensive search.

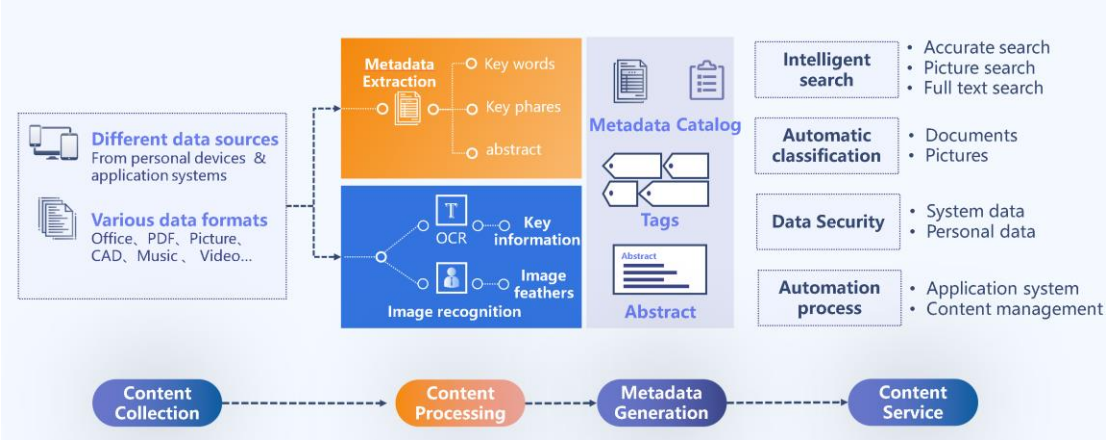

Figure 3 The Dada management of IHEP Docs

2.2Data Security Module

IHEP Docs establishes a security technology architecture and a standard management process to protect data from data transfer, platform capabilities, and application control. Data security module (Figure 4), including content security, access security, data protection, storage security and so on, provides reliable security management and control capabilities to ensure a whole life-cycle data security including the process of data production/collection, data transmission, data storage, data processing, data exchange, data archiving/destruction, etc., and avoids the leakage and loss of data. Meanwhile, an online anti-virus and anti-ransomware module was deployed and integrated with IHEP Docs, and will perform real-time virus scanning when documents are uploaded.

IHEP Docs are deployed in IHEP and all the data are stored in a storage cluster fully managed by IHEP. IHEP Docs data are stored in three copies, and each document is placed on a different node and in a different hard disk. Once a hard disk or a node is damaged, IHEP Docs will not be affected and the data will still be safe. A more secure transmission technology, HTTPS, was adopted based on TLS protocol to ensure data transmission security. IHEP Docs also provides an open backup API to perform scheduled incremental backups of specified document libraries, and restore data at any time once the system or data are damaged. The maintenance permission of IHEP Docs is separated from the document access permission, which can effectively protect users' private data.

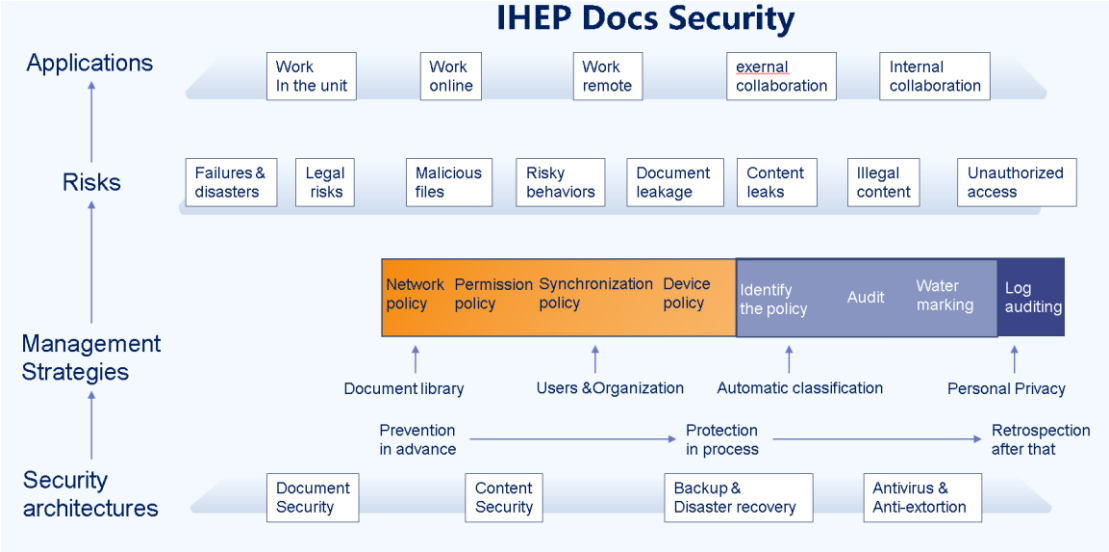

Figure 4 Data security architecture of IHEP Docs

2.3System Deployment

IHEP Docs was deployed in the computing center of IHEP with a privatization multi-machine cluster mode. Based on different functions, the system divides different service modules which can be deployed independently and the main module adopts cloud-native microservice mode for multi-machine deployment (Figure 5). Each server has two 10 Gigabit fibers for the splitting and integration of document content and two 10 Gigabit fibers for reception of the request and response to ensure a quick respond to the data requests.

A distributed storage and a scale-out hardware expansion mode are adopted in a cluster mode for distributed data storage and paralleled data processing. Once a node fails, the service can automatically switch to other normal nodes to ensure sustainable access to services. IHEP Docs uses an OSS gateway to form multiple object storages into a heterogeneous multi-object storage cluster and provide a unified storage access and independently manage multi-object storage. The databases of MariaDB and MongoDB are deployed in a highly available manner to avoiding the effect of a single point failure.

The data of IHEP Docs are saved in multiple copies with the CRUSH algorithm to ensure different copies distributed on different nodes, which can avoid performance bottlenecks caused by the imbalance loads of different node, and reduce the risk of data corruption caused by node or disk failure. The system has a high reliability based on Linux virtual server + keepalive and container technology [13] and can automatically detect node failures and restore affected containers on other available nodes by monitoring container status and automatic recovery.

IHEP Docs provides a unified workspace and open API interfaces for changing data and integrating applications, such as WPS Office Online and CAD online, and deploys a high-performance storage cluster on the back-end to provide high-performance and high-throughput data services.

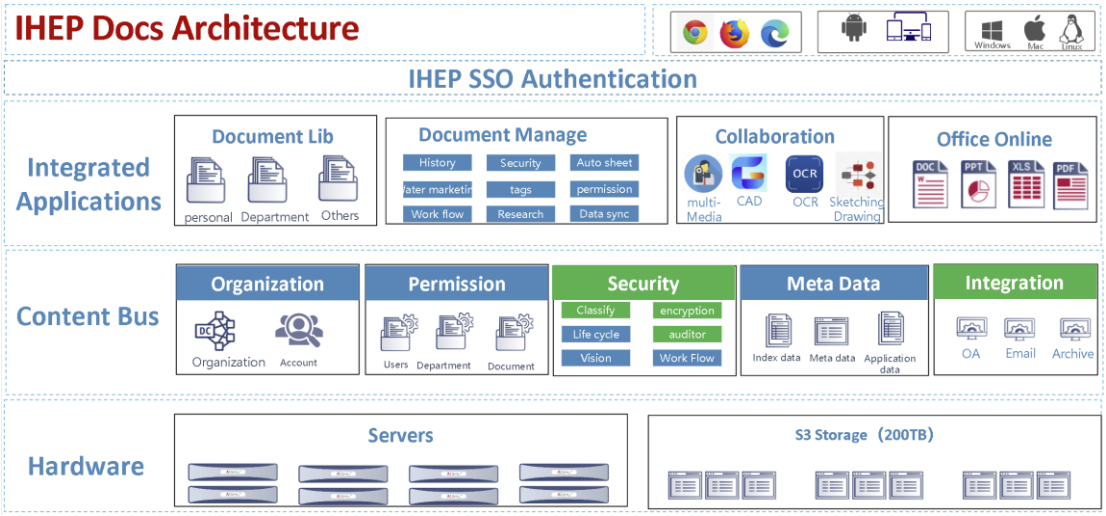

Figure 5 The architecture and modules of IHEP Docs

3. System Functions and Applications

3.1System Functions

IHEP Docs is an open document management system based on content bus service, and can be used to manage various unstructured data with a full life cycle process, such as classification, release, collaboration, retrieval, utilization, archiving, and destruction. It provides users a unified workspace with consistent experience and various application service such as data generation, data collection, business processes, office collaboration, data management, and data consumption.

3.1.1Open Content Bus

The content bus (Figure 6) provides various APIs, such as RESTful API, OAuth 2.0, and content integration open framework, etc. The data collected from third-party application by OpenDoc API, the data collected from IHEP Docs applications by Open APIs, and the data collected from personal terminal devices by Client/APP/Web browser can be managed in a unified workspace and used for circulation, query, management, retrieval, and multi-terminal access.

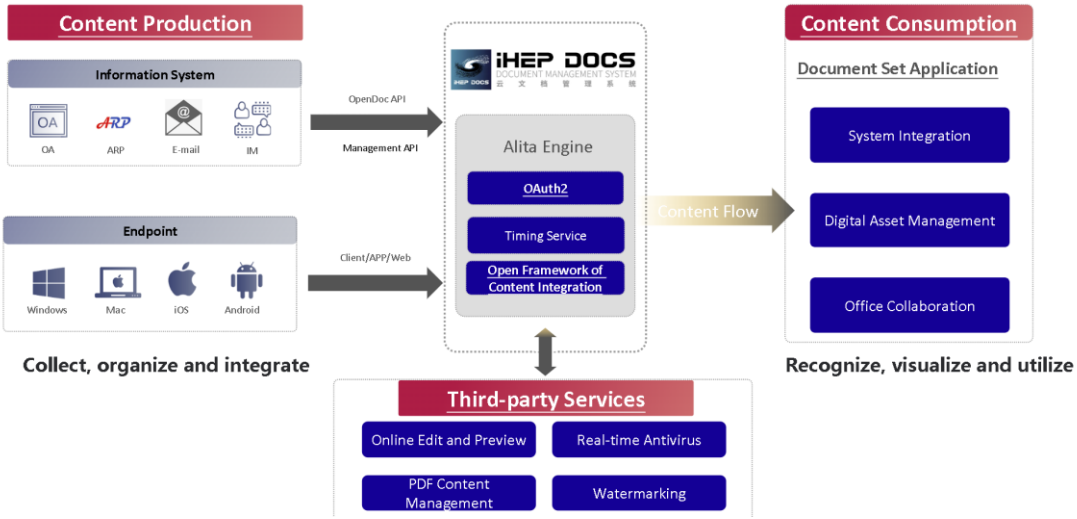

Figure 6 Content bus architecture of IHEP Docs

3.1.2Open Integrated development framework

IHEP Docs has an integrated development framework (Figure 7), which provides third-party applications a series of open API interfaces. A redevelopment based the content bus APIs, or a targeted plug-ins based on the Web Widget framework can be achieved for personalized requires. After authorized, the third-party applications can access the data resources of IHEP Docs, break out the system boundaries and data boundaries between application systems, and utilize the related application data by the open API interfaces. IHEP Docs has integrate many powerful applications such as WPS office online, antivirus, CAD, etc. and provides users a unified workspace.

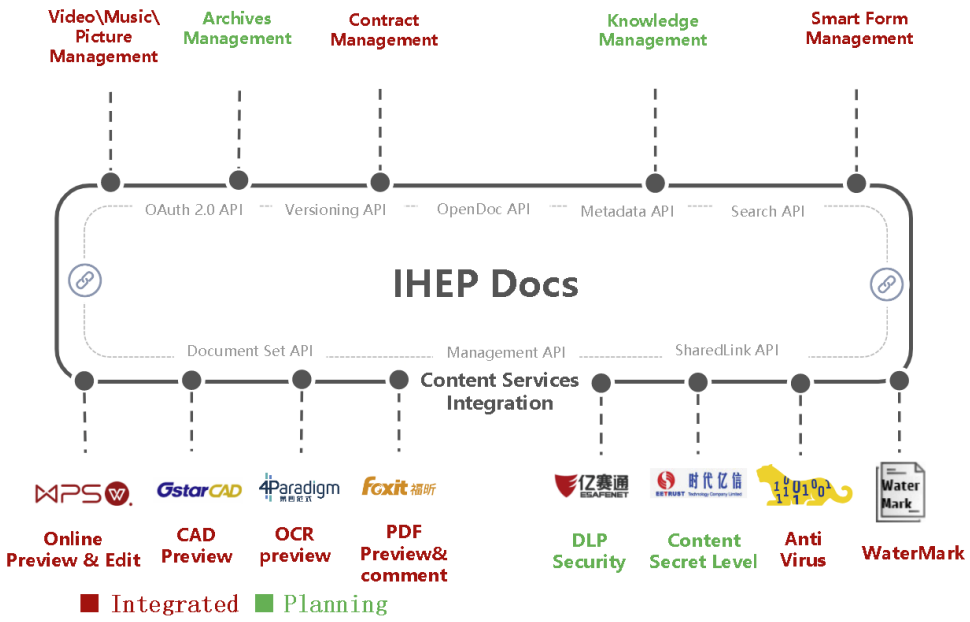

Figure 7 Integrated development framework of IHEP Docs

3.1.3Controllable Permission Management

IHEP Docs supports unified file operations in the workspace and fine-grained permissions such as display, preview, download, create, modify, delete, reject, owner permission, etc. (figure 8). The access, sharing, or use of documents can be strictly restricted by permission control to ensure document security. Mobile devices can be managed and the data can be erased from remote once the mobile device is lost. Meanwhile, IHEP Docs supports an access control based on networks, devices, or document libraries by yourself, so that the right people have the right access to the right files.

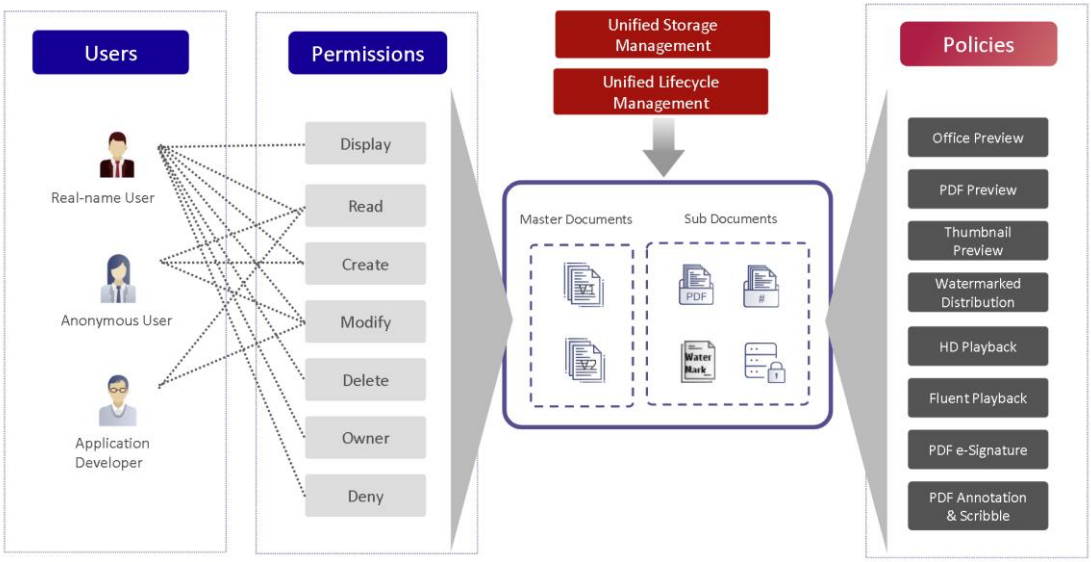

Figure 8 Permission management of IHEP Docs

3.2Integrated Applications

IHEP Docs has unified storage, unified standards, unified interfaces, unified data management and unified data utilization. The unstructured data collected are stored and unified in a centralized manner, the third-party applications are integrated by an open APIs, the data are transferred in unified standards, consistent services are provided in a unified workspace.

Oriented to the management of the whole life cycle process of unstructured data in the process of science, technique, and management, IHEP Docs breaks down data islands and form a unified data asset, links different application system through content bus service to realize the orderly flow of unstructured data, realize the efficient connection between content management and integrated applications. It is promising to eliminate data islands, reduce the cost of management and operation, realize the centralized storage, unify the data management, and office collaboration.

IHEP Docs can help to effectively manage the data throughout the life cycle and ensure the data security. According to different document management requires of departments and large-scale scientific facilities in IHEP, IHEP Docs is configured with different document libraries to gather and manage scattered and heterogeneous unstructured data. Integrated with anti-virus, IHEP Docs can give a real time protection when documents uploaded or downloaded to protect the data security, realize the effective document management, and empower digital applications in IHEP.

3.2.1Unified Workspace and Document Management

IHEP Docs provides a unified workspace with document management and content service for users, realize unified authentication by integration with IHEP SSO (Figure 9). It also provides powerful integrated applications in a consistent and powerful use experience on different devices. Users can work together in the workspace, share documents, preview online, edit document online, comment online, collect files, and audit the files.

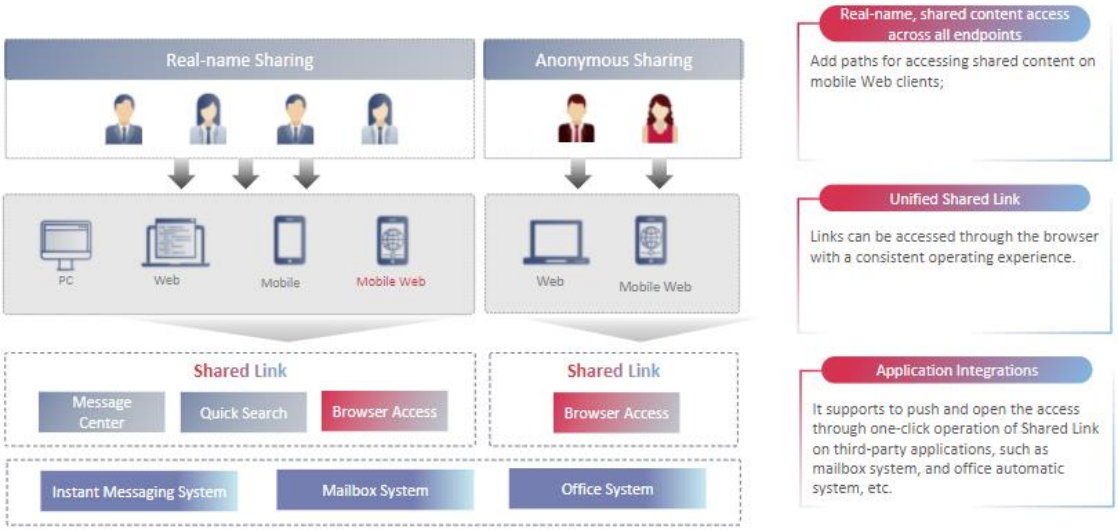

Figure 9 Workspace with other integrated applications

3.2.2Collaboration Office Online

Integrated with powerful applications, IHEP Docs provides more efficient office document collaborations for content sharing, document approval, document archive, document editing together, and automatic data collection. IHEP Docs integrates WPS office online, CAD online and other office collaborative software, and provide users an online collaborative platform to preview, edit, revise, and share documents with others online at the same time to reduce the communication cost effectively during documents being edited by many people (Figure 10). All modified or edited documents, as a history version, can be reviewed or restored anytime.

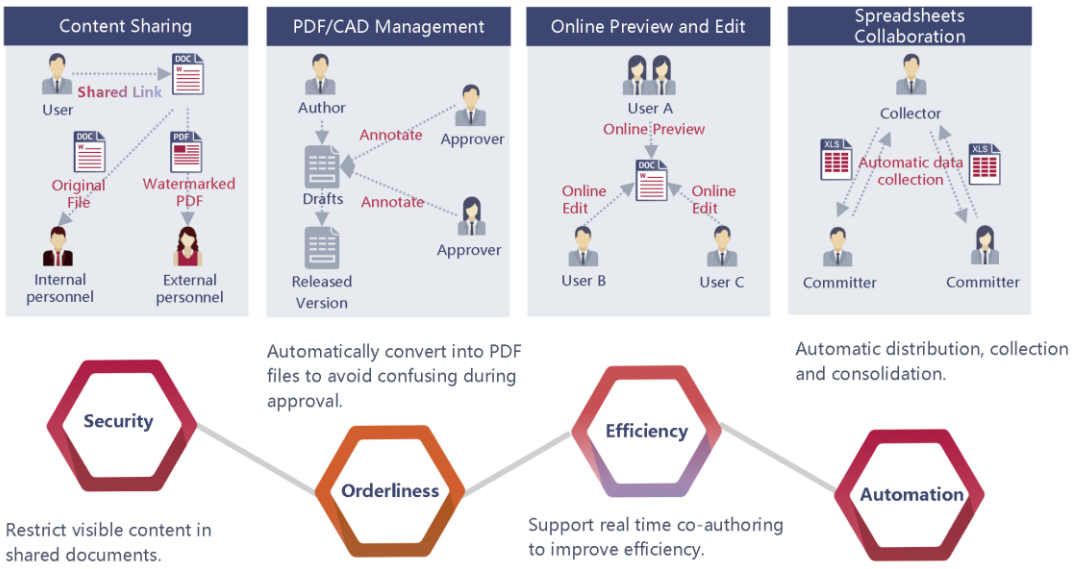

Figure 10 Office collaboration online of IHEP Docs

3.2.3Intelligent search

By massive index and metadata combined with OCR and deep learning, IHEP Docs provides a fast search, full-text search, tag search and image search (Figure 11). When the various format documents are collected from applications and devices, the system will recognize the documents or pictures and generate the metadata and tags. These analyzed metadata are extracted and

classified for easy search. When users search a word or a picture, the related documents and pictures will be found out and listed.

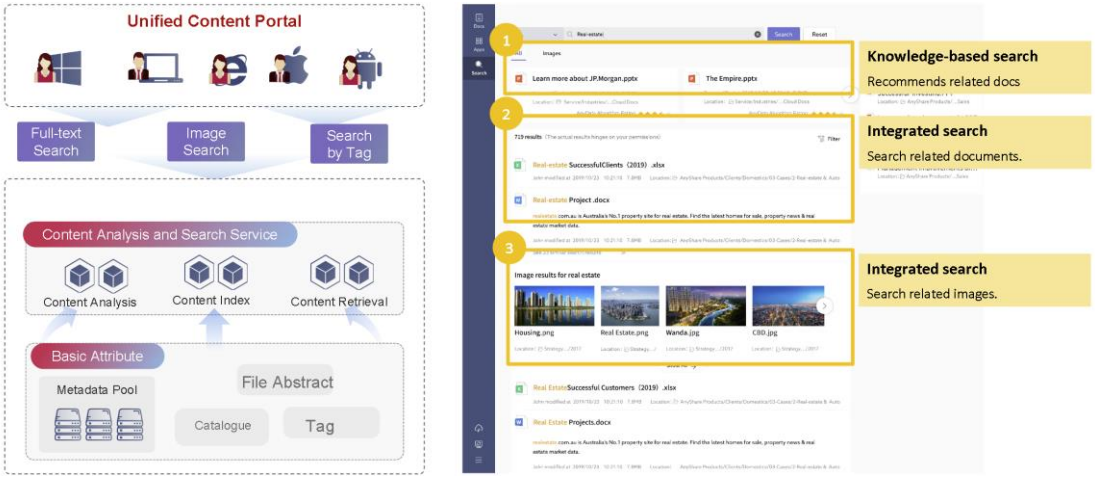

Figure 11 Intelligent search of IHEP Docs

3.2.4Document Work Flows

IHEP Docs integrates document work flows according to document management and document collection, which can be customed by users to review, submit or gather documents (Figure 12). User can create a document review workflow, and set several auditors. when a document is summitted, the auditors will be noticed. After audited, the document will be sent to some specified directory and saved.

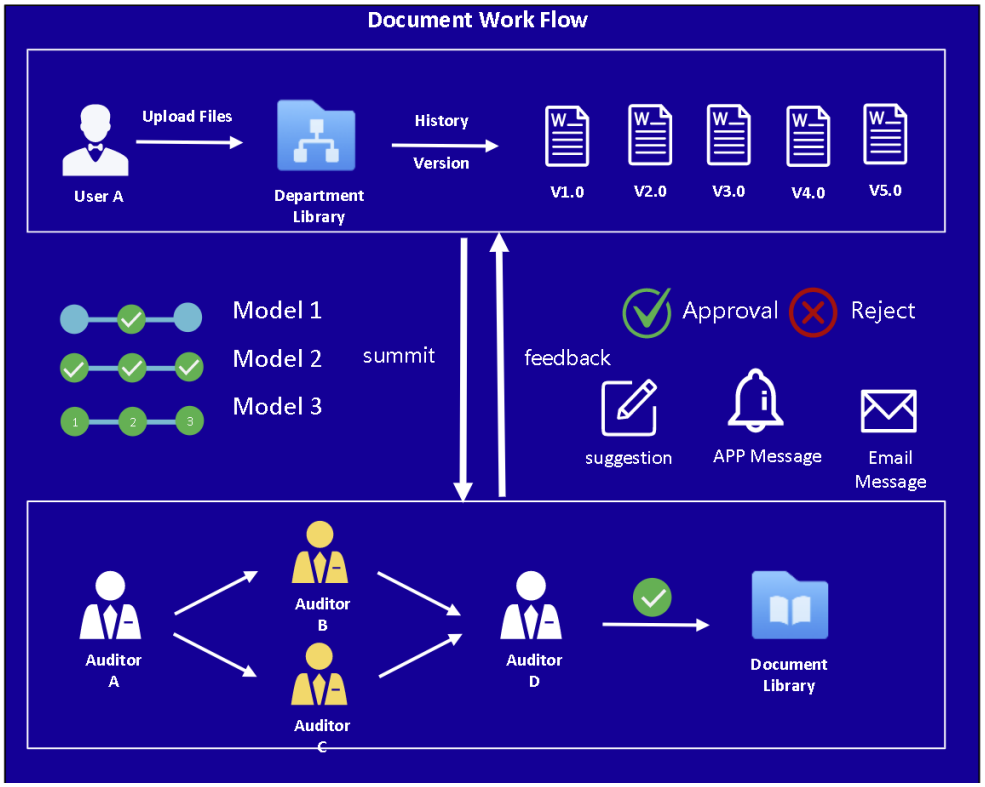

Figure 12 Document work flow of IHEP Docs

3.2.5Smart Collaborative Form

IHEP Docs integrates a smart collaborative form, by which users can quickly create forms to share with others by a link or a QR code and collect data filled by others (Figure 13). Submitters can only see the form for them to fill in. It can organize different sheets into standardized one automatically.

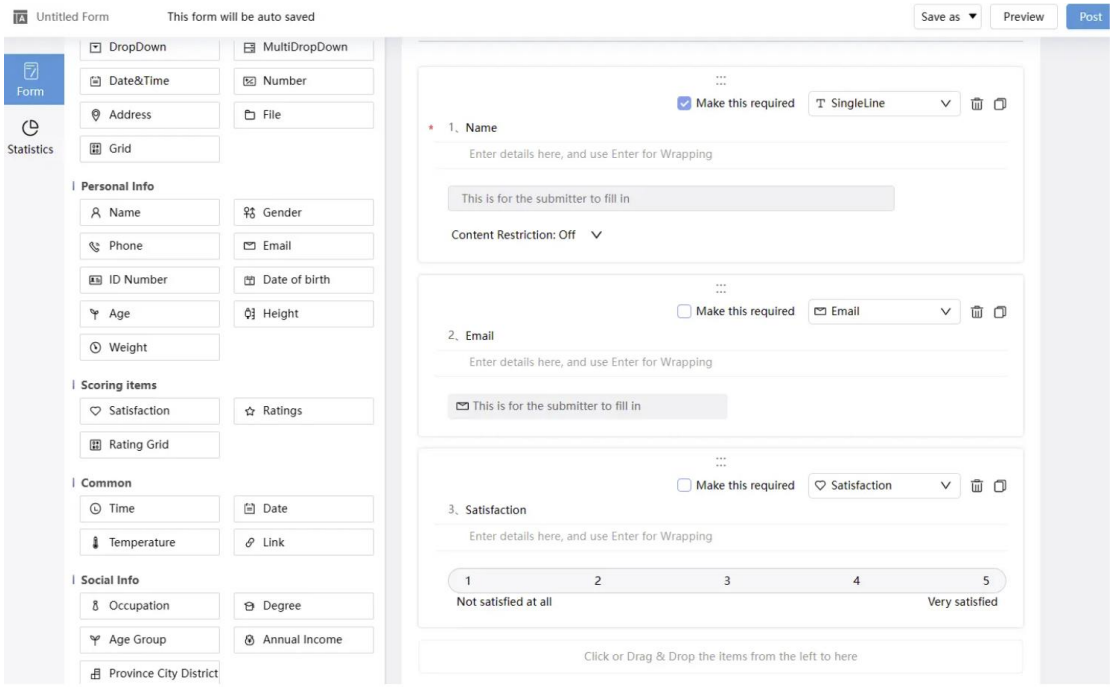

Figure 13 Smart collaboration form of IHEP Docs

4. System operation

IHEP Docs system has been online for almost 8 months since September 2022 and it is adopted by almost all the departments and more than 20 large-scale scientific facilities. Thousands of libraries have been created to manage document and edit files together online. There are more than 200 accounts active every day, and the user signed in and use IHEP Docs from all over the world.

There are over 380,000 documents stored in IHEP Docs, which occupies about 3 terabytes of storage space, and the distribution of file types is shown in figure 15. The cumulative number of user operations on files has reached more than 2.5 million and the operation type distribution is shown in figure 16. Up to now, IHEP Docs has been disinfected about 140,000 times and removed more than 200 viruses contained in files, ensuring the data security.

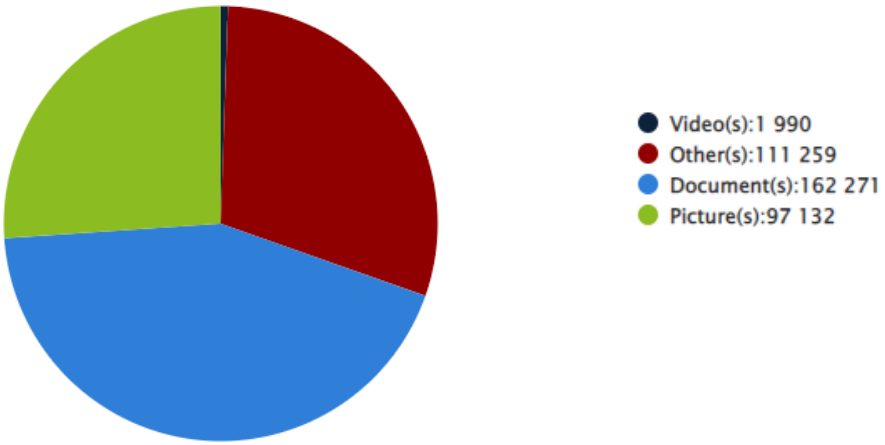

Figure 14 Document distribution of different file types

5. Conclusion and Outlook

For the requirements of document management and document collaboration, we developed and deployed IHEP Docs. At present, IHEP Docs have been widely used in IHEP. As planned, we will base on the specific requirements in document management, develop IHEP Docs and integrate the existing information systems in IHEP to achieve the orderly aggregation, efficient management and deep mining of unstructured data stored in these systems.

In order to realize further improvement in data management and knowledge service finally, there are more difficult jobs to be done in future to gather, sort, analyze, process those unstructured data, form a knowledge service that can be retrieved and reused easily, such as knowledge communities, wiki, and Q&A. At the same time, it is promising to develop a computable intelligent form to integrate work flow and ultimately provide users with powerful, easy-to-custom data service. The system security and data security are further strengthened by automatic encryption and decryption technology to build up a more security system.

6. Acknowledgements

This work was supported by a grant from the National Natural Science Foundation of China (nos. 11905239), Youth Innovation Promotion Association CAS (292021000091), and Cyber security and informatization projects CAS(CAS-WX2021SF-0406).

References

[1] A. Doan, J. F. Naughton, R. Ramakrishnan, et al., *Information Extraction Challenges in Managing Unstructured Data, SIGMOD record: ACM SIGMOD* **37** (4) 14-20.

[2] A. Gandomi, M. Haider, *Beyond the hype: big data concepts, methods and analytics, International Journal of Information Management* **35** (2) 137-144.

[3] J. Liao, *Research on Information Extraction and Fusion of Knowledge Graph for Unstructured Data, University of Electronic Science and Technology of China* **2021**.

[4] Z. Yao, H. Cao, *The Management Strategy of the Unstructured Information in MES, Applied Mechanics and Materials* **48–49** 1271–1274.

[5] J. Yang, Y. Liu, T. Qi, *Documents Datafication: Concept, Framework and Methods. Journal of Library Science in China* **48** (259) 063-078.

[6] BESIII Collaboration, *The construction of the BESIII experiment, Nuclear Instruments and Methods in Physics Research, Section A. Accelerators, Spectrometers, Detectors and Associated Equipment* **598** (1) 7-11.

[7] S. Wang, S. Fang, S. Fu, *Introduction to the overall physics design of CSNS accelerators, Chinese Physics C* **33** (z2) 1-3.

[8] Y. Tao, *Ground breaking Ceremony at the High Energy Photon Source in Beijing[J]. Synchrotron Radiation New* **32** (5) 40.

[9] G. Ren, T. Wang, D. Wang, *Construction and Application of Interactive Assessment Environment Based on Document Cloud, 2021 Tenth International Conference of Educational Innovation through Technology (EITT)* **2021** 234-239.

[10] M. Gupta, A. Choudhary, J. Parmar, *Analysis of Text Identification Techniques Using Scene Text and Optical Character Recognition, International journal of computer vision and image processing* **11** (4) 39-62.

[11] K. Vukatana, *OCR and Levenshtein distance as a measure of image quality accuracy for identification documents, 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)* **2022** 1-4.

[12] N. Azginoğlu, M. A. Eren, M. Çelık, et al., *Ceph-based storage server application, 2018 6th International Symposium on Digital Forensic and Security (ISDFS)* **2018** 1-4.

[13] B. David, *Containers and Cloud: From LXC to Docker to Kubernetes, IEEE cloud computing* **1** (3) 81-84.