International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Y.3604

(02/2020)

SERIES Y: GLOBAL INFORMATION INFRASTRUCTURE, INTERNET PROTOCOL ASPECTS, NEXT-GENERATION NETWORKS, INTERNET OF THINGS AND SMART CITIES

Cloud Computing

Big data – Overview and requirements for data preservation

Recommendation ITU-T Y.3604

International Telecommunication Union

ITU-T Y-SERIES RECOMMENDATIONS

GLOBAL INFORMATION INFRASTRUCTURE, INTERNET PROTOCOL ASPECTS, NEXT-GENERATION NETWORKS, INTERNET OF THINGS AND SMART CITIES

GLOBAL INFORMATION INFRASTRUCTURE	
General	Y.100–Y.199
Services, applications and middleware	Y.200–Y.299
Network aspects	Y.300–Y.399
Interfaces and protocols	Y.400–Y.499
Numbering, addressing and naming	Y.500–Y.599
Operation, administration and maintenance	Y.600–Y.699
Security	Y.700–Y.799
Performances	Y.800–Y.899
INTERNET PROTOCOL ASPECTS	
General	Y.1000–Y.1099
Services and applications	Y.1100–Y.1199
Architecture, access, network capabilities and resource management	Y.1200–Y.1299
Transport	Y.1300–Y.1399
Interworking	Y.1400–Y.1499
Quality of service and network performance	Y.1500–Y.1599
Signalling	Y.1600–Y.1699
Operation, administration and maintenance	Y.1700–Y.1799
Charging	Y.1800–Y.1899
IPTV over NGN	Y.1900–Y.1999
NEXT GENERATION NETWORKS	
Frameworks and functional architecture models	Y.2000–Y.2099
Quality of Service and performance	Y.2100–Y.2199
Service aspects: Service capabilities and service architecture	Y.2200–Y.2249
Service aspects: Interoperability of services and networks in NGN	Y.2250–Y.2299
Enhancements to NGN	Y.2300–Y.2399
Network management	Y.2400–Y.2499
Network control architectures and protocols	Y.2500–Y.2599
Packet-based Networks	Y.2600–Y.2699
Security	Y.2700–Y.2799
Generalized mobility	Y.2800–Y.2899
Carrier grade open environment	Y.2900–Y.2999
FUTURE NETWORKS	Y.3000–Y.3499
CLOUD COMPUTING	**Y.3500–Y.3999**
INTERNET OF THINGS AND SMART CITIES AND COMMUNITIES	
General	Y.4000–Y.4049
Definitions and terminologies	Y.4050–Y.4099
Requirements and use cases	Y.4100–Y.4249
Infrastructure, connectivity and networks	Y.4250–Y.4399
Frameworks, architectures and protocols	Y.4400–Y.4549
Services, applications, computation and data processing	Y.4550–Y.4699
Management, control and performance	Y.4700–Y.4799
Identification and security	Y.4800–Y.4899
Evaluation and assessment	Y.4900–Y.4999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T Y.3604

Big data – Overview and requirements for data preservation

Summary

Recommendation ITU-T Y.3604 provides the overview of big data preservation and its requirements which are derived from the corresponding use cases. It addresses the subjects of overview of big data preservation, functional requirements of big data preservation as well as use cases of big data preservation.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T Y.3604	2020-02-06	13	11.1002/1000/14138

Keywords

Big data, challenge, data category, data package, data preservation, functional entity, functional requirement, overview, strategy.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T Y.3604

Big data – Overview and requirements for data preservation

1 Scope

This Recommendation provides overview and requirements of big data preservation. It addresses the following subjects:

– Overview of big data preservation;

– Functional requirements of big data preservation;

– Use cases of big data preservation.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T X.1601] Recommendation ITU-T X.1601 (2015), *Security framework for cloud computing*.

[ITU-T Y.3600] Recommendation ITU-T Y.3600 (2015), *Big data – Cloud computing based requirements and capabilities*.

[ITU-T Y.3603] Recommendation ITU-T Y.3603 (2019), *Big data – Requirements and conceptual model of metadata for data catalogue*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following term defined elsewhere:

3.1.1 big data [ITU-T Y.3600]: A paradigm for enabling the collection, storage, management, analysis and visualization, potentially under real-time constraints, of extensive datasets with heterogeneous characteristics.

NOTE – Examples of datasets characteristics include high-volume, high-velocity, high-variety, etc.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 big data preservation: Combination of policies, strategies and actions developed by a big data service provider (BDSP) to ensure that digital information of continuing value remains accessible and usable within a big data ecosystem.

NOTE – A big data ecosystem defines necessary activities for roles providing and consuming big data services as well as relationships between roles (see [ITU-T Y.3600]).

3.2.2 data preservation: The policies and actions to ensure continued access to data.

3.2.3 data preservation policy: A set of rules for controlling the data preservation actions.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

BDSP Big Data Service Provider

DAP Data Access Package

DPP Data Preserved Package

DSP Data Selection Package

5 Conventions

In this Recommendation:

The keywords "**is required to**" indicate a requirement which must be strictly followed and from which no deviation is permitted if conformance to this document is to be claimed.

The keywords "**is recommended**" indicate a requirement which is recommended but which is not absolutely required. Thus this requirement need not be present to claim conformance.

6 Overview of big data preservation

This clause presents an overview of big data preservation. It describes the general concept of data preservation and illustrates the data preservation in big data ecosystem. It also describes data categories, data packages and functional entities of big data preservation.

6.1 General concept of data preservation

Data preservation is the policies and actions to ensure continued access to data. Maintaining the safety and integrity of data and its metadata for further use is the fundamental principle of data preservation. The goal of data preservation is to protect data from being lost or destroyed and to render data accurately over time.

Data can be lost or destroyed in many different ways, including the deterioration and ageing of storage media, the disappearance of readable hardware or software and the impossibility of reading the format of the files containing the data, etc.

Data preservation exceeds the concept of having or possessing data or back-up copies of data. Data preservation supports persistent access to data by planning back-up and recovery strategies, preceding the event of a disaster or technological changes.

6.2 Data preservation in big data ecosystem

When applied to a big data ecosystem, data preservation typically involves the combination of policies, strategies and actions to ensure that digital information of continuing value remains accessible and usable within the big data service provider (BDSP).

In the following, the term "big data preservation "is used to refer to "data preservation "in a big data ecosystem".

6.2.1 Challenges for big data preservation

The characteristics of big data (see clause 6.1 of [ITU-T Y.3600]) bring challenges for big data preservation as follow:

– **volume**: refers to the amount of data collected, stored, analysed and visualized, which big data technologies need to resolve. This characteristic brings challenges for:
  - how to scale the data storage; and
  - how to determine which data to preserve.

NOTE – Data that are not used directly may provide important value by indirect means, like analytics. Therefore, it can be more complicated to decide what to preserve and when to take action to preserve large volumes of data.

– **variety**: refers to different data types and data formats that are processed by big data technologies. Each type of data may need a hardware and software environment to render. This characteristic brings challenges for:
  • how to preserve both media and their execution context.
– **velocity**: refers to both how fast the data are being collected and how fast the data are being processed by big data technologies to deliver expected results. For big data preservation, it means that:
  • preservation activity should become more automated.

6.2.2 Strategies for big data preservation

According to [ITU-T Y.3600], a BDSP supports data management activities such as data provenance, data privacy, data security, data retention policy, data ownership, etc. Although not explicit listed, data preservation is also one of data management activities and it offers continued access to data according to a data preservation policy, which is a set of rules for controlling the data preservation actions.

For overcoming the above challenges, a BDSP may adopt the following strategies for big data preservation:

– **automated operations**: the operations for big data preservation include selecting and preparing the data to be preserved, preserving and retrieving the preserved data. Because of the huge volume and high velocity of the big data, a BDSP should keep these operations and linkages among them "automatic" for higher efficiency;
– **supporting flexible data preservation policy**: big data preservation should be a policy-driven process. Which data should be preserved, which type of storage media should be used and how long the preserved data should be kept are all examples of issues to be determined by flexible data preservation policies. Policy-based big data preservation can help a BDSP overcoming the challenges for volume and velocity;
– **easy scaling and hierarchical data storage**: huge volumes of data require more data storage. Considering the balance of economy and efficiency of the preservation data storage, a BDSP should employ a hierarchical data storage system and each tier of this system should be easy to scale out.

6.3 Data categories of big data preservation

The following are different categories of data in big data preservation:

– **content data**: refers to the raw data to be preserved;
– **metadata**: refers to data about data or data elements, possibly including their data descriptions, and data about data ownership, access paths, access rights and data volatility. The general concept of metadata and its utilization on a big data ecosystem have been described in [ITU-T Y.3603]. For supporting big data preservation, following types of information are needed:
  • **representation information**: describes the structure and semantics of the content data (see clause 8.2.3 of [ITU-T Y.3603]). It also includes other information that may be needed to understand the content data, e.g., software and algorithms, etc.;
  • **reference information**: provides access information and finding aids (e.g., subject category and keywords) of the content data (see clause 8.2.3 of [ITU-T Y.3603]);

- **provenance information**: documents the history of the content data (see clause 8.2.6 of [ITU-T Y.3603]);
- **fixity information**: provides the data integrity checks to protect the content data from undocumented alteration;
- **access rights information**: provides the terms of access to the content data (see clause 8.2.3 of [ITU-T Y.3603]).

– **rule data**: refers to the rules set by a BDSP and should be referenced and obeyed when executing activities of big data preservation. Representative rules for big data preservation include preservation selection rules which specify what kind of data should be preserved in the BDSP, preservation period rules which specify how long data should be preserved in the BDSP and preservation storage rules which specify which storage tier these data should be stored in and in which conditions they should be migrated to other storage tiers, etc.

NOTE – For example, to automate the process of selecting data to be preserved, the BDSP can set a preservation selection rule to specify what kind of data should be preserved. The parameter of this rule could be related with data type, data size and keywords in filename, etc.

6.4 Data packages within big data preservation

Data packages contain content data and related metadata. There are three types of data packages within big data preservation:

1) **data selection package (DSP)**: contains the data to be preserved and related metadata (normally including representation information and access rights information), provided by data selection management (see clause 6.5);
2) **data preserved package (DPP)**: contains the preserved data and metadata, generated and maintained by preserved data storage management (see clause 6.5);
3) **data access package (DAP)**: contains the requested preserved data and related metadata (normally including representation information), provided by data access management (see clause 6.5).

The relationship between DPP and DAP could be one-to-one, one-to-many, many-to-one, many-to-many. For example, according to different request conditions, the data of a DAP could come from one DPP or many DPPs. The relationship between DPP and DSP is as complex as relationship between DPP and DAP.

6.5 Functional entities for big data preservation management

The functional entities for managing big data preservation consists of preservation policy management, data selection management, preserved data storage management and data access management, as shown in Figure 6-1. Functional entities are as follows:

– **preservation policy management**: manages preservation policy to ensure that the preserved data remains accessible and usable;

– **data selection management**: supports selecting data to be preserved and preparing the DSP for storage according to relevant rules;

– **preserved data storage management**: supports storing the data to be preserved and maintaining the preserved data including refreshing and migration according to relevant rules;

– **data access management**: supports handling access requests for the preserved data and providing the requested data to the requester by DAP.

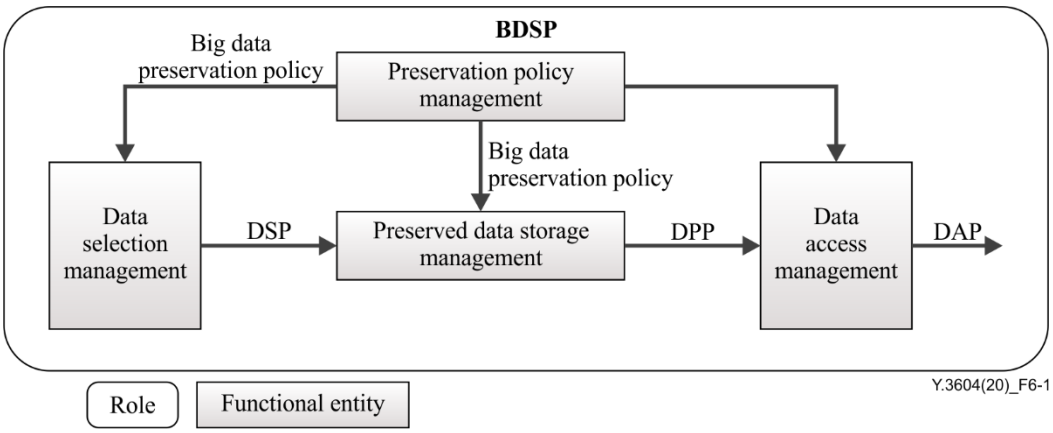


Figure 6-1 – Functional entities for big data preservation management

NOTE – Figure 6-1 represents one role of a BDSP.

7 Functional requirements of big data preservation

This clause identifies functional requirements applicable to big data preservation.

7.1 Selecting the data to be preserved

The selecting the data to be preserved requirements include:

– it is recommended that a BDSP supports selecting the data to be preserved from the original data according to the preservation selection rules;

NOTE 1 – The original data includes all of the data collected from outside of the BDSP and processing results generated inside the BDSP.

NOTE 2 – The preservation selection rules specify what kind of data should be preserved in the BDSP. The parameter of these rules could be related with data type, data size and keywords in filename, etc.

– It is required that a BDSP supports extracting the metadata from the data to be preserved;

NOTE 3 – The metadata extracted from the data to be preserved normally includes representation information, fixity information and access rights information.

– It is required that a BDSP supports generating the DSP which is composed of content data and metadata.

7.2 Storing the data to be preserved

Storing the data to be preserved requirements include:

– it is required that a BDSP supports converting DSP into DPP;

– it is recommended that a BDSP supports selecting the proper storage tier to store the DPP according to the preservation storage rules;

NOTE 1 – The preservation storage rules specify which storage tier these data should be stored in and in which conditions they should be migrated to other storage tiers.

– it is recommended that a BDSP supports monitoring the DPP's access statistics to manage its storage hierarchy;

– it is required that a BDSP supports removing the preserved data according to the preservation period rules.

NOTE 2 – The preservation period rules specify how long the data should be preserved in a BDSP.

7.3 Accessing the preserved data

The accessing the preserved data requirements include:

– it is required that a BDSP supports verifying privileges of the preserved data requestor to access the preserved data;

– it is required that a BDSP supports searching preserved database on filter conditions provided by the preserved data requestor;

– it is required that a BDSP supports converting DPP into DAP.

7.4 Managing the data preservation policy

The managing the data preservation policy requirements include:

– it is recommended that a BDSP supports setting preservation selection rules to specify what kind of data should be preserved in a BDSP;

– it is recommended that a BDSP supports setting preservation period rules to specify how long these data should be preserved in a BDSP;

– it is recommended that a BDSP supports setting preservation storage rules to specify which storage tier these data should be stored in and in which conditions they should be migrated to other storage tiers.

8 Security considerations

Relevant security requirements of [b-ITU-T Y.2201], [b-ITU-T Y.2701] and applicable X, Y and M series of ITU-T Recommendations need to be taken into consideration, including access control, authentication, data confidentiality, data retention policy, network security, data integrity, availability and protection of personal information.

Appendix I

Use cases of big data preservation

(This appendix does not form an integral part of this Recommendation.)

I.1 Use case template

The use cases developed in Appendix I should adopt the following unified format for better readability and convenient material organization.

Table I.1 – Use case template

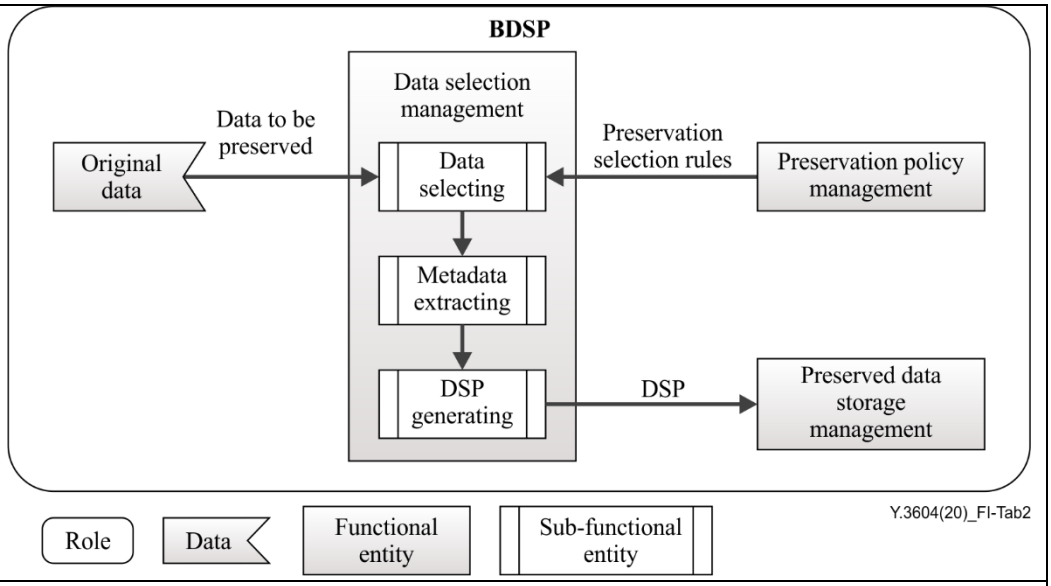
Title	Note: The title of the use case.
Description	Note: Scenario description of the use case.
Roles	Note: Roles involved in the use case.
Figure (optional)	Note: Figure to explain the use case, but not mandatory.
Pre-conditions (optional)	Note: The necessary pre-conditions that should be achieved before starting the use case.
Post-conditions (optional)	Note: The post-condition that will be carried out after the termination of current use case.
Derived requirements	Note: Requirements derived from the use cases, whose detailed description is presented in the dedicated chapter.

I.2 Use case of selecting the data to be preserved

Table I.2 – Selecting the data to be preserved

Title	Selecting the data to be preserved
Description	To prepare the data to be preserved, the BDSP selects the data based on the following procedures: – the BDSP selects the data to be preserved (i.e. content data) from the original data according to the preservation selection rules which specify what kind of data should be preserved in the BDSP; NOTE 1 – The original data includes all of the data collected from outside of the BDSP and processing results generated inside the BDSP. NOTE 2 – The preservation selection rules specify what kind of data should be preserved in the BDSP. The parameter of these rules could be related with data type, data size and keywords in filename, etc. – the BDSP extracts some of the metadata (e.g., representation information, fixity information and access rights information) from the data to be preserved; – the BDSP generates the DSP which is composed of content data and metadata.
Roles	BDSP
Figure (optional)	

Table I.2 – Selecting the data to be preserved

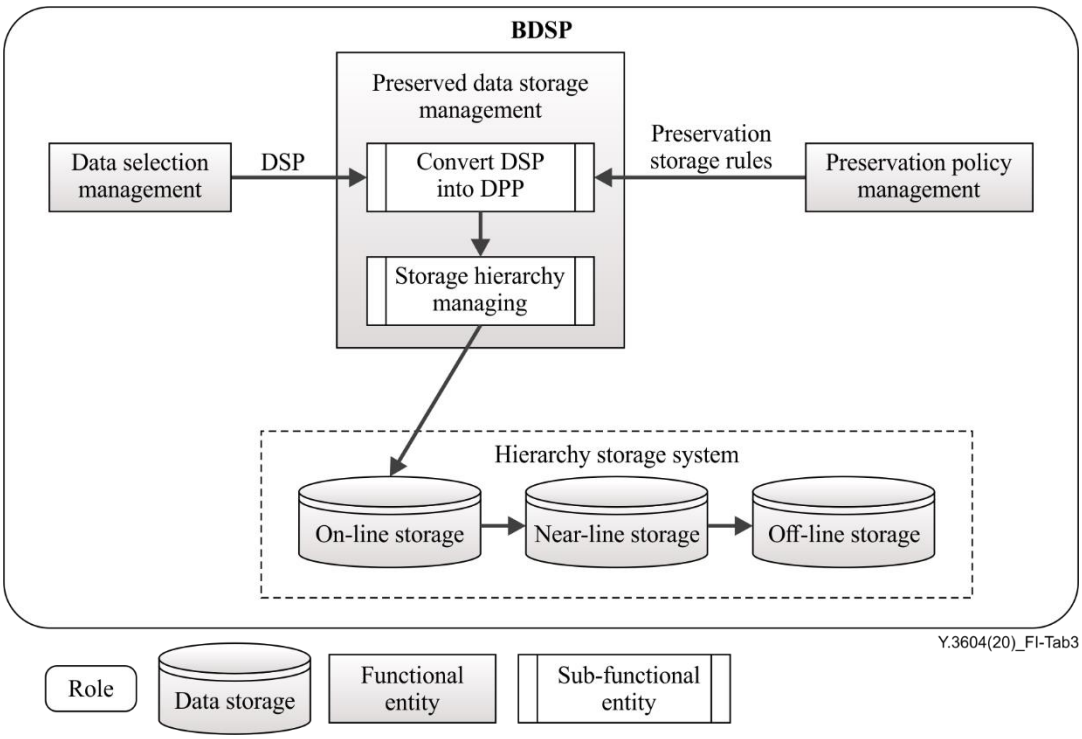
	
Pre-conditions (optional)	The preservation selection rules have been set by the BDSP.
Post-conditions (optional)	
Derived requirements	Select the data to be preserved (refer to clause 7.1).

I.3 Use case of tiering preserved data storage

Table I.3 – Tiering preserved data storage

Title	Tiering preserved data storage
Description	For storing the data to be preserved, a BDSP has a hierarchical storage system which has three tiers: on-line storage, which means the data are immediately available; near-line storage, which means the data are not immediately accessible but is reasonably quickly accessible; and off-line storage, which means that the data are totally disconnected from the network and requires the most time to recover. How to select a proper storage tier to store the data to be preserved is managed by the storage hierarchy managing function and based on the preservation storage rules. For example, there is a preservation storage rule specifying that if the size of the content data is less than 10 GB, it should be stored in on-line storage, otherwise it should be stored in near-line storage directly. And if the access times to these data within 3 months are less than 3 times, it should be moved to a lower storage tier, e.g., from on-line storage to near-line storage. A DSP which contains 6 GB of content data will be stored by the following procedures: – the BDSP converts the DSP into DPP; – according to the preservation storage rules described above, the DPP should be stored in the on-line storage; – the BDSP will keep monitoring the DPP's access statistics to determine whether it should be subsequently moved to lower storage tier after 3 months; – the BDSP will keep checking according to the preservation storage rule when the preserved data should be removed from the BDSP.
Roles	BDSP

Table I.3 – Tiering preserved data storage

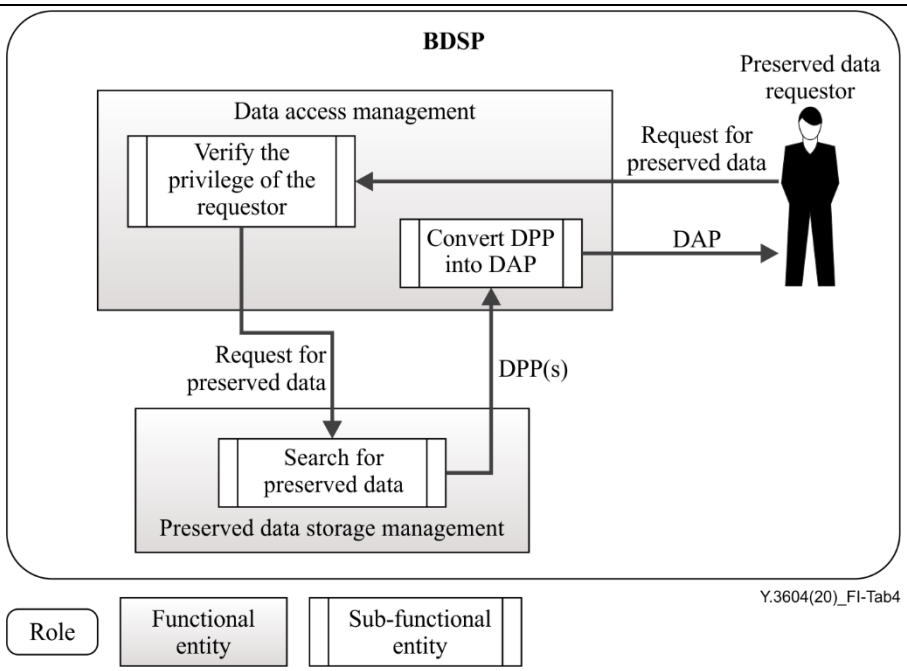
Figure (optional)	
Pre-conditions (optional)	The preservation storage rules have been set by the BDSP. A DSP has been provided by data ingestion management to preserved data storage management.
Post-conditions (optional)	The DPP has been stored in a proper storage tier according to the data preservation rule.
Derived requirements	Storing the data to be preserved (refer to clause 7.2).

I.4 Use case of accessing preserved data

Table I.4 – Accessing preserved data

Title	Accessing preserved data
Description	The preserved data requestor wants to access some preserved data from the BDSP. The preserved data will be accessed based on the following procedures: – the preserved data requestor sends a request with filter conditions to the data access management to specify what data it wants to access; – the data access management verifies if the preserved data requestor has the privilege to access the requested data, and then sends a request to preserved data storage management for getting the preserved data; – the preserved data storage management searches for the requested preserved data according to the filter conditions and sends the resulting DPP(s) to the data access management; – the data access management converts the DPP(s) to DAP and sends it back to the preserved data requestor.
Roles	BDSP

Table I.4 – Accessing preserved data

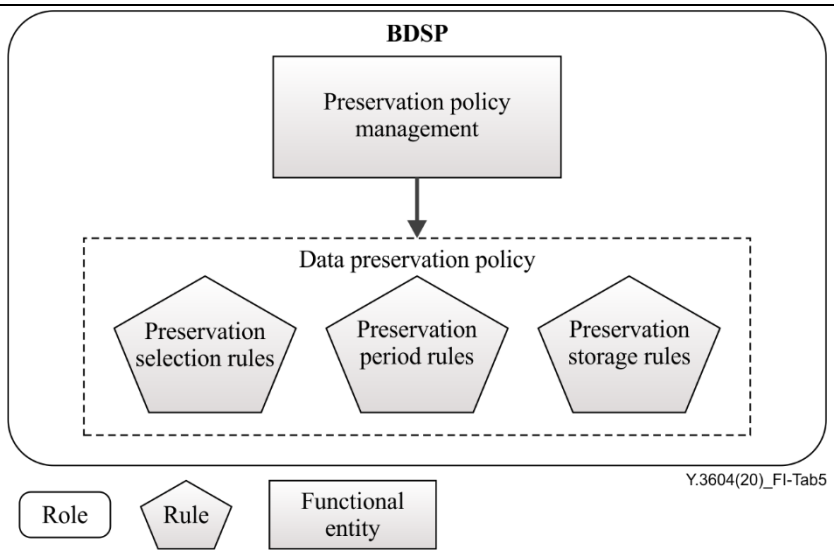
Figure (optional)	
Pre-conditions (optional)	The data to be accessed has been preserved in the BDSP.
Post-conditions (optional)	
Derived requirements	Accessing the preserved data (refer to clause 7.3).

I.5 Use case of data preservation policy management

Table I.5 – Data preservation policy management

Title	Data preservation policy management
Description	The BDSP implements the data preservation management based on the data preservation policy. The BDSP can set data preservation policy through preservation planning management. The data preservation policy consists of: – **preservation selection rules**: specify what kind of data should be preserved in the BDSP. The parameter of this rule could be related with data type, data size and keywords in filename, etc.; – **preservation period rules**: specify how long these data should be preserved in the BDSP; – **preservation storage rules**: specify which storage tier these data should be stored in and in which conditions they should be migrated to other storage tiers.
Roles	BDSP

Table I.5 – Data preservation policy management

Figure (optional)	
Pre-conditions (optional)	
Post-conditions (optional)	
Derived requirements	Managing the data preservation policy (refer to clause 7.4).

Bibliography

[b-ITU-T Y.2201] Recommendation ITU-T Y.2201 (2009), *Requirements and capabilities for ITU-T NGN*.

[b-ITU-T Y.2701] Recommendation ITU-T Y.2701 (2007), *Security requirements for NGN release 1*.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	**Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities**
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2020