

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Y.3603

(12/2019)

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

Cloud Computing

Big data – Requirements and conceptual model of metadata for data catalogue

Recommendation ITU-T Y.3603

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

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.3603

Big data – Requirements and conceptual model of metadata for data catalogue

Summary

Recommendation ITU-T Y.3603 describes the general concept of metadata and its utilization in a big data ecosystem. Also, this Recommendation provides requirements and a conceptual model of metadata for data catalogue as well as the extensible markup language (XML) schema of metadata as an example. This metadata supports finding data easier, and is used for exchange, preservation, integration and provenance of data in a big data ecosystem.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T Y.3603	2019-12-14	13	11.1002/1000/14137

Keywords

Big data, big data catalogue, conceptual model, metadata for big data, metadata requirements.

* 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/>.

© ITU 2020

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

		Page
1	Scope.....	1
2	References.....	1
3	Definitions	1
	3.1 Terms defined elsewhere	1
	3.2 Terms defined in this Recommendation.....	1
4	Abbreviations and acronyms	1
5	Conventions	2
6	Overview of metadata.....	2
	6.1 Introduction to metadata in the big data ecosystem	2
	6.2 Introduction to big data lifecycle.....	3
7	Requirements of metadata for data catalogue.....	4
	7.1 Metadata for data catalogue.....	4
	7.2 Requirements of metadata elements	5
8	Conceptual model for data catalogue	6
	8.1 Overall class diagram	6
	8.2 Classes and data type	7
	Appendix I – Examples of metadata profile for data catalogue.....	13
	I.1 RDF profile.....	13
	I.2 Turtle profile.....	21
	Appendix II – Examples of metadata instance.....	27
	II.1 UML object diagram	27
	II.2 RDF instance	27
	Bibliography.....	38

Recommendation ITU-T Y.3603

Big data – Requirements and conceptual model of metadata for data catalogue

1 Scope

This Recommendation introduces the metadata concept as well as its usages in a big data lifecycle. This Recommendation provides requirements and a conceptual model of metadata for a data catalogue to support the big data ecosystem defined in [ITU-T Y.3600]. The extensible markup language (XML) profiles based on the conceptual model described by the unified modelling language (UML) and examples of metadata are provided in the appendices.

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 Y.3600] Recommendation ITU-T Y.3600 (2015), *Big data – Cloud computing based requirements and capabilities*.
- [ITU-T Y.3601] Recommendation ITU-T Y.3601 (2018), *Big data – Framework and requirements for data exchange*.
- [ITU-T Y.3602] Recommendation ITU-T Y.3602 (2018), *Big data – Functional requirements for data provenance*.
- [ITU-T Z.109] Recommendation ITU-T Z.109 (2016), *Specification and Description Language – Unified modeling language profile for SDL-2010*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 data catalogue [ITU-T Y.3601]: A listing of all metadata which a data broker makes available.

3.1.2 metadata [b-ITU-T H.752]: Structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment and management of the described entities.

3.2 Terms defined in this Recommendation

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

- API Application Programming Interface
- BDC Big Data Service Customer

BDSP	Big Data Service Provider
DCAT	Data Catalogue Vocabulary
DP	Data Provider
JSON	JavaScript Object Notation
RDF	Resource Description Framework
Turtle	Terse Resource Description Framework Triple Language
UML	Unified Modelling Language
URI	Uniform Resource Identifier
XML	extensible Markup Language

5 Conventions

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.

The keywords "**can optionally**" indicate an optional requirement which is permissible, without implying any sense of being recommended. This term is not intended to imply that the vendor's implementation must provide the option and the feature can be optionally enabled by the network operator/service provider. Rather, it means the vendor may optionally provide the feature and still claim conformance with the specification.

In the body of this document and its annexes, the words shall, shall not, should, and may sometimes appear, in which case they are to be interpreted, respectively, as is required to, is prohibited from, is recommended, and can optionally. The appearance of such phrases or keywords in an appendix or in material explicitly marked as informative are to be interpreted as having no normative intent.

The unified modelling language (UML) schema used in this Recommendation follows [ITU-T Z.109].

6 Overview of metadata

6.1 Introduction to metadata in the big data ecosystem

Metadata is structured, encoded data that describe characteristics of a dataset to aid in the identification, discovery, assessment and management of the dataset. Metadata summarizes overall information about data, making finding and working with particular data easier. Typically, the metadata structure is highly dependent on data format or application. Therefore, a metadata for big data is needed to provide an overarching model to cover the characteristics of big data.

Figure 6-1 illustrates metadata in the big data ecosystem defined in [ITU-T Y.3600]. Metadata is used for data publication and search among roles and sub-roles of the big data ecosystem.

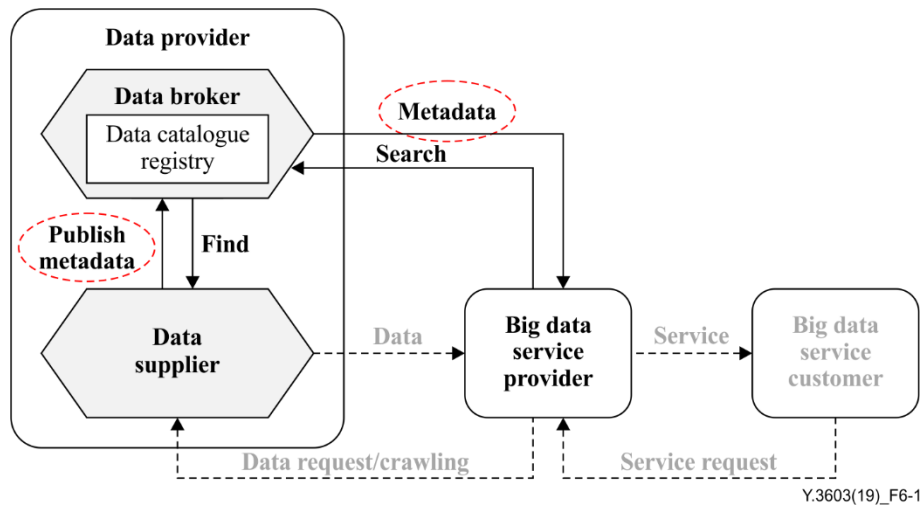


Figure 6-1 – Metadata in the big data ecosystem

Regarding the data provider (DP), DP:data broker provides a meta-information registry (e.g., data catalogue registry) to DP:data suppliers for publishing their data sources, and DP:data suppliers are creating and publishing metadata information to DP:data broker. A big data service provider (BDSP) searches information through metadata stored in the DP:data broker's data catalogue registry.

According to [ITU-T Y.3600], metadata provides:

- critical information for identification, access right of data;
- information about creation, control, and update on data for the management purpose;
- additional information about data integration capabilities.

6.2 Introduction to big data lifecycle

Metadata is used for various purposes, such as preserving data, data distribution, data integration and data history management by continuously recoding and updating. Figure 6-2 shows the usages of metadata according to big data lifecycle.

NOTE – An example of big data lifecycle has a main cycle with creation/capture, storage, preparation, analysis, visualization, and also has branch cycles with retirement/deletion and share/exchange.

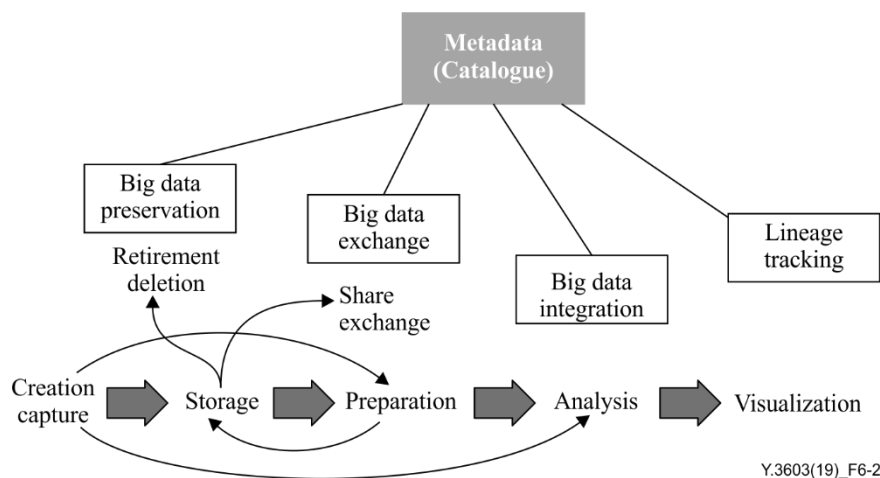


Figure 6-2 – Usage of metadata in big data lifecycle

[ITU-T Y.3601] defines the functional requirements for big data exchange. In that Recommendation, DP:data broker provides a common data catalogue schema to exchange big data which covers various

type of data sources. DP:data broker provides meta-information (data catalogue for dataset) registry to data source. A data source publishes metadata to an online registry for data sharing. This metadata includes:

- data type, format, size and its update frequency;
- category (area of data source);
- metadata and its access information (e.g., uniform resource identifier (URI));
- delivery mechanism, and electronic access methods to data (e.g., application programming interface (API)).

And, this metadata can optionally include:

- data rights (e.g. general agreement, copyrights, patents, licenses policies and pricing policies) (e.g. the applied functions with parameter);
- usage of data;
- quality of data (e.g., sales ranking, and data supplier's credit);
- related keywords; and
- sample data.

DP:data broker also supports data classification with commonly used data vocabulary and taxonomy, and it is recommended to support multiple data classifications by usage area.

[ITU-T Y.3602] defines the functional requirements for big data provenance. Provenance is information about the origin and creation process of data, including the metadata recording process of workflow, annotations and notes about processes. When data are imported from an outside data source and stored, the BDSP generates metadata based on the importing context, including:

- responsible party information;
- time stamp;
- size, etc.

The metadata can be used for recording provenance information according to [ITU-T Y.3602].

For storage efficiency and performance, the BDSP provides data preservation capabilities by monitoring information (e.g., time stamp) included in metadata. In addition, the BDSP uses metadata to retrieve the data for data preparation (e.g. data integration).

7 Requirements of metadata for data catalogue

7.1 Metadata for data catalogue

A data catalogue describes the requirements for searching a dataset. It helps the users of the data to evaluate the dataset to determine its suitability of usage of the data. This clause provides six elements and contents of a data catalogue for big data. Figure 7-1 shows the elements of metadata for a data catalogue.

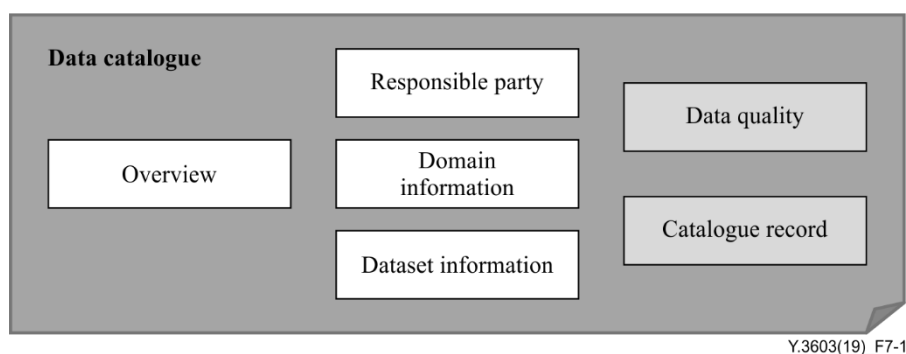


Figure 7-1 – Elements of metadata for a data catalogue

A data catalogue **is required to** include the following elements:

- overview (see clause 7.2.1);
- responsible party (see clause 7.2.2);
- dataset information (see clause 7.2.3);
- domain information (see clause 7.2.4).

A data catalogue **is recommended to** include the following elements:

- data quality (see clause 7.2.5);
- catalogue record (see clause 7.2.6).

7.2 Requirements of metadata elements

7.2.1 Overview element

The overview element provides overall information of the data catalogue and is required to include:

- information about the creation of the data catalogue;
NOTE 1 – This may include the title, the language used in the text, and the version of data catalogue specification.
- an informal description of the data catalogue.
NOTE 2 – An informal description is intended to give a short introduction to data catalogue and allow a human reader for better understanding.

7.2.2 Responsible party element

The responsible party element provides authoritative reference information and is required to include:

- information to identify and communicate with person(s) and organizations associated with the data catalogue.
NOTE 1 – This may include the role or function performed by the organization, the organization name, individual name and position name.

Responsible party element is recommended to include following:

- contact information.
NOTE 2 – This may include the phone number, address, online resource such as URI and time period when individuals may contact the publisher of data catalogue.

7.2.3 Dataset information element

A data catalogue element contains one or more datasets. The dataset information of metadata represents the actual dataset as published by the DP:data provider, and is required to include:

- the title of the dataset;

- an informal description of the dataset.

NOTE 1 – An informal description is intended to give a short introduction to the dataset and allow a human reader a better understanding.

The dataset information element is recommended to include:

- dataset distribution information;

NOTE 2 – A dataset may have multiple distribution information.

NOTE 3 – This information may include the distribute format of dataset, type of dataset (structured, semi-structured, unstructured), access information including data API, user's guide or its URI, size of dataset, sample data for user's preview and data rights.

- keywords representing dataset;
- contact point information.

NOTE 4 – Contact point information is represented with the same contents of the responsible party described in clause 7.2.2, but its target is DP:data provider.

Dataset information may also include the following information optionally:

- the frequency at which the dataset is updated;
- temporal coverage of dataset which is described by period of time.

7.2.4 Domain information element

The domain information element provides the application area of dataset and is required to include:

- source domain of the dataset for classification.

The domain information element is recommended to include:

- usage domain of the dataset.

7.2.5 Data quality element

In a big data ecosystem, there are various types and formats of data, each having different quality requirements. Therefore, different quality evaluation methods are applied to different datasets. The data quality element is required to include:

- quality requirement on the dataset;
- unit of measure for data quality;
- quality evaluation result.

7.2.6 Catalogue record element

The catalogue record element provides date information for registration and modification for instances of the data catalogue. The catalogue record element is required to include:

- the most recent modified date of data catalogue.

The catalogue record element of data catalogue is recommended to include:

- the issued date of listing the dataset in the catalogue.

NOTE – The issued date of listing a dataset in a data catalogue is different than the publication date of the dataset.

8 Conceptual model for data catalogue

8.1 Overall class diagram

This clause provides a UML model based on clause 7, which describes the requirements of metadata for a data catalogue. Figure 8-1 shows the overall class diagram of the data catalogue using UML to provide the design of a concept for a data catalogue.

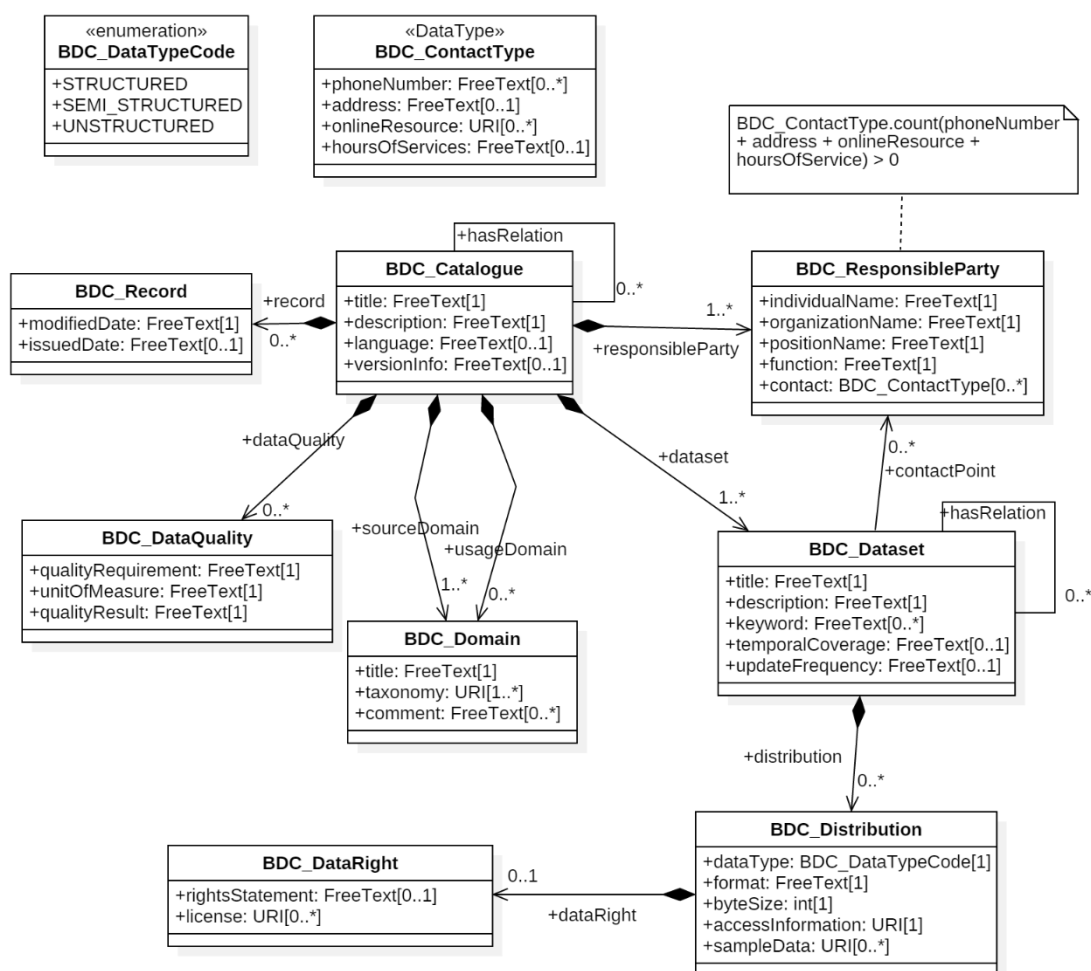


Figure 8-1 – Overall class diagram of data catalogue using UML

8.2 Classes and data type

The following descriptors are used to explain the data dictionaries in the clauses that follow.

- Name/Role name: title of class (gray colored row in table), association (noted as role) and attribute;
- Description: short description about 'Name/Role name';
- M/R/O: the mandatory/recommended/optional (M/R/O) necessity of each class, association and attribute;
- Cardinality: the maximum number of instances that the metadata entity or metadata element may have. Single occurrences are shown by "1"; repeating occurrences are represented by "N".
- Domain: the values allowed, or the use of free text placed on the contents of the field. "Free text" indicates that no restrictions are placed on the contents of the field. Integer-based codes shall be used to represent values for domains containing codelists.
- Domain code: designated code used in place of the concept.

8.2.1 Class: BDC_Catalogue

Table 8-1 shows the data dictionary of *BDC_Catalogue* which describes the overview element information in clause 7.2.1.

Table 8-1 – Catalogue overview

	Name/Role name	Description	M/R/O	Cardinality	Domain
1	BDC_Catalogue	root entity which defines metadata about dataset resource	M	1	Line 2 to 12
2	role: responsibleParty	responsible party information for the data catalogue	M	N	BDC_ResponsibleParty
3	role: dataset	dataset information for the data catalogue	M	N	BDC_Dataset
4	role: sourceDomain	source domain information for the data catalogue	M	N	BDC_Domain
5	role: usageDomain	usage domain information for the data catalogue	R	N	BDC_Domain
6	role: dataQuality	data quality information for the data catalogue	R	N	BDC_DataQuality
7	role: record	record information for the data catalogue	R	N	BDC_Record
8	role: hasRelation	related data catalogue of the data catalogue	O	N	BDC_Dataset
9	title	title of data catalogue	M	1	free text
10	description	a short introduction to data catalogue	M	1	free text
11	language	language used within the data catalogue	R	1	free text (e.g., RFC 5646 [b-RFC5646])
12	versionInfo	version of data catalogue specification	R	1	free text (major.minor.patch)

8.2.2 Class: BDC_ResponsibleParty

Table 8-2 shows the data dictionary of *BDC_ResponsibleParty* and its data type class *BDC_ContactType* which describes the responsible party in clause 7.2.2.

Table 8-2 – Responsible party information

	Name/Role name	Description	M/R/O	Cardinality	Domain
13	BDC_ResponsibleParty	name and position information for an individual or organization that is responsible for the data catalogue or dataset	M	N	Line 14 to 18
14	individualName	name of the responsible person: surname, given name, title	M	1	free text
15	organisationName	name of the responsible organization	M	1	free text
16	positionName	role or position of the responsible person	M	1	free text