RDA Data Foundation and Terminology - DFT: Results RFC

Gary Berg-Cross, Raphael Ritz, Peter Wittenburg,

State: 29.6.2015
Version 1.5

The goals of the DFT Working Group were:

- Moving the discussion in the data community towards an agreed, upon suite of terms relevant to RDA groups with a focus on basic core model of related terms along with some basic principles that will harmonize the data organization solutions.
- Fostering an effective RDA community culture by converging on essential terminology arising from agreed upon reference models.

Based on a variety of data models and use cases presented by experts coming from different disciplines and about 120 interviews and interactions with different scientists and scientific departments, the DFT WG has composed a number of simple definitions for digital data in a registered[1] domain based on group conceptualization and synthesis.

1. DFT Core Term Definitions

1.1 Digital Object (DO)

Definition

A digital object (DO) is represented by a bitstream, is referenced and identified[2] by a persistent identifier and has properties that are described by metadata.

Note: As indicated we only talk about registered DOs in the context of this document.
Note: Properties included in metadata include discovery, contextual, schema, rights, curation and provenance information.
Note: A DO is said to be a dynamic DO when the information content represented in a DO is changing for some period of time or even for an indefinite duration.

1.2 Persistent Identifier (PID)

Definition

A persistent identifier is a long-lasting ID represented by a string that uniquely identifies a DO and that is intended to be persistently resolved to meaningful state information about the identified DO[3].
Note: We use the term Persistent Resolvable Identifier as a synonym.

[1] There will always exist data in private, temporary stores, which will not be organized and made accessible in a standard way.

[2] Various repositories include passport-like metadata information along with the PID which goes beyond pure referencing.

[3] This can be information such as checksum, access paths, references to additional information, etc. Some repositories call this administrative or system metadata, but some think a minimal set of attributes to be included has not been well defined yet. In cases where the digital objects do not exist anymore, due to finite lifetime for example, the PID is expected to continue to exist and can still be resolved into useful information.

1.3 PID Record

Definition

A PID record contains a set of attributes stored with a PID describing DO properties.

1.4 PID Resolver (aka Resolution System)

Definition

A PID resolution system is a globally available infrastructure system that has the capability to resolve a PID into useful, current state information describing the properties of a DO[4].

1.5 Metadata

Definition

Metadata contains descriptive, contextual and provenance assertions about the properties of a DO.

Note: Such metadata will make the DO for example discoverable, accessible and usable/interpretable.

Note: To make metadata referable it needs to be associated with a PID and thus is a DO.

Note: Metadata minimally needs to contain the PID of the DO.

1.6 Aggregation

Definition

A digital aggregation is a bundle of digital entities.

Note: The term "aggregation" as a base concept does not add substantially to our understanding of the intuitive idea of collections as resulting from some aggregation process and thus is not used as a separately defined concept.

1.7 Digital Collection

Definition

A digital collection is an aggregation which contains DOs and DEs. The collection is identified by a PID and described by metadata.

Note: A digital collection is a (complex) DO.

Note: A digital collection is an aggregation in so far as there are other types of aggregations.

1.8 Digital Entity

Definition

A digital entity is anything that can be represented by a bitstream.

1.9 Repository

Definition

A digital repository is an infrastructure component that is able to store, manage and curate DOs and return their bitstreams when a request is being issued.

1.10 Bitstream

Definition

A bitstream is a sequence of bits that encodes a specific content, either stored on some media or being transferred under control of protocols.

Note: The term "bit-sequence" is seen as a synonym in the context of DFT.

1.11 State Information

Definition

State information is "metadata" information that describes those current properties of the DO that are relevant for proper management and access.

[4] There are a couple of comparisons such as http://www.clarin.eu/content/comparison-pid-systems

1.12Property

Definition

A property of a digital object specifies one of its characteristics as digital data.

1.13 Metadata Repository

Definition

A digital metadata repository is a digital repository that is able to store, manage and curate metadata.

Note: A metadata repository is a digital repository.

Note: Metadata can be aggregated by service providers to registries or catalogues.

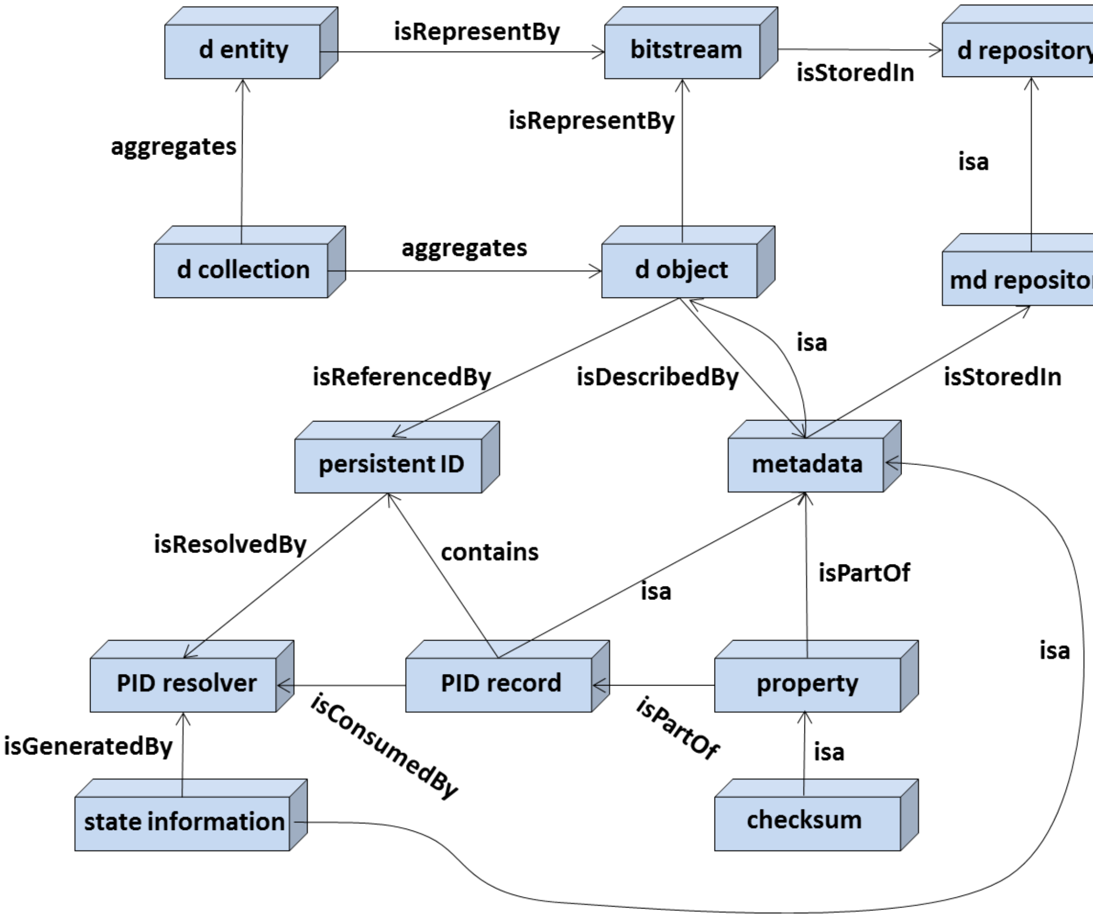
1.14 Checksum

Definition

A checksum is metadata and an important property of a digital object to allow verifying identity and integrity.

2. Basic Data Organization model

The data organization model which is the basis for the above term definitions has been drawn from the models that have been suggested to overcome the current deficits in dealing with data. It should be noted that this model only applies to the domain of registered data and that there may always be special requirements in the area of very large data sets for example.



3. Associated Documents

Major document products of the DFT WG were:

- **DFT 1: Model Overview**
  An annotated collection of data organization & management models that represent concrete use cases, i.e. models that are foundational to running data systems or that specific communities associated with RDA are considering using as the basis of their data systems.

- **DFT 2: Analysis & Synthesis**
  In this document we analyzed the models overviewed in part 1. Two major, complementary model categories were noted: ones describing ***data organizations*** (describing a model) and others focused more on the processing of data according to certain ***workflows***. Analytic summaries of each are provided followed by a synthesis which employs a common conceptualization, depicted graphically, to draw a number of conclusions.

- **DFT 3: Term Snapshot**
  An overview of some core terms and their relations capturing as the DFT WG wrapped up its efforts. Methods and consolidated, core definitions are reviewed based in analysis and synthesis discussed in other documents. The intent of the core snapshot is to be used subsequently as a platform to accelerate discussions towards real, working agreements on terminology within RDA and across the worldwide data community. As an aid to this an Appendix on examples of data management is provided.

- **DFT 4: Use Cases**
  A collection of use case scenarios developed by the community and discussed at Plenaries as examples of relevant work. These use pertinent term concepts such as PID, Digital Object or Research Data Object. In addition graphics are presented along with additional textual propositions that assert what we should capture in our definitions or issues with concepts. On the whole these use cases focus on repository operations such as registration of PIDs as supported by PID systems.

- **DFT 5: Term Tool Description**
  This document described the RDA DFT WG Term Definition Tool (aka TeD-T) -a web application for collecting and discussing term definitions. The application is freely available for read access and after a free registration, users are also able to edit existing content or create new entries. Existing terms can be browsed as alphabetical list or structured by the underlying hierarchy of TeD-T, a fulltext search completes the browsing functionality. The document includes figures to explain how the tool works and tutorial is available on the site. The TeD-T platform is deployed and maintained at RZG.