Data Quality

Angelica Lo Duca
IIT-CNR
angelica.loduca@iit.cnr.it

Linked Open Data:
a paradigm for the
Semantic Web

Definition

- Data Quality is *fitness for usage**
- Data is only good as its quality
- Affects potential use of the data
  - Limits full exploitation of such data by data consumers

** Juran J. The Quality Control Handbook. McGraw -Hill, New York, 1974*

Data Quality Assessment

- Data Quality assessment involves measuring of quality ***dimensions*** *or criteria* that are relevant to the consumer

Quality Dimensions

- Accessibility
- Intrinsic
- Contextual
- Representational

Accessibility

- Availability
- Licensing
- Interlinking
- Security
- Performance

Availability is the extent to which information (or some portion of it) is present, obtainable and ready for use.

Availability metrics

- A1: checking whether the server responds to a SPARQL query
- A2: checking whether an RDF dump is provided and can be downloaded
- A3: detection of dereferenceability of URIs by checking:
  - for dead or broken links, i.e. that when an HTTP-GET request is sent, the status code 404 Not Found is not returned
  - that useful data (particularly RDF) is returned after lookup of a URI
  - for changes in the URI, i.e. compliance with the recommended way of implementing redirections using the status code 303 See Other
- A4: detect whether the HTTP response contains the header field stating the appropriate content type of the returned file, e.g. application/rdf+xml
- A5: dereferenceability of all forward links: all available triples where the local URI is mentioned in the subject (i.e. the description of the resource)

Licensing is the granting of permission for a consumer to re-use a dataset under defined conditions.

Licensing metrics

- L1: machine-readable indication of a license in the VoID description or in the dataset itself
- L2: human-readable indication of a license in the documentation of the dataset
- L3: detection of whether the dataset is attributed under the same license as the original

Interlinking is the degree to which entities that represent the same concept are linked to each other, be it within or between two or more data sources.

Interlinking metrics

- I1: detection of:
  - interlinking degree: how many hubs there are in a network
  - clustering coefficient: how dense is the network
  - centrality: indicates the likelihood of a node being on the shortest path between two other nodes
  - whether there are open sameAs chains in the network
  - how much value is added to the description of a resource through the use of sameAs edges
- I2: detection of the existence and usage of external URIs (e.g. using owl:sameAs links)
- I3: detection of all local in-links or back-links: all triples from a dataset that have the resource's URI as the object

Security is extent to which data is protected against alteration and misuse.

Security metrics

- S1: using digital signatures to sign documents containing an RDF serialization, a SPARQL result set or signing an RDF graph
- S2: verifying authenticity of the dataset based on provenance information such as the author and his contributors, the publisher of the data and its sources (if present in the dataset)

Performance is the efficiency of a system that binds to a large dataset, that is, the more performant a data source the more efficiently a system can process data.

Performance metrics

- P1: checking for usage of slash-URIs where large amounts of data is provided
- P2: low latency - (minimum) delay between submission of a request by the user and reception of the response from the system
- P3: high throughput - (maximum) number of answered HTTP-requests per second
- P4: scalability: detection of whether the time to answer an amount of ten requests divided by ten is not longer than the time it takes to answer one request

Quality Dimensions

- Accessibility
- **Intrinsic**
- Contextual
- Representational

Intrinsic dimensions

- Semantic accuracy
- Syntactic validity
- Consistency
- Conciseness
- Completeness

Syntactic validity is the degree to which an RDF document conforms to the specification of the serialization format.

Syntactic validity metrics

- SV1: detecting syntax errors using validators or via crowdsourcing
- SV2: detecting use of:
  - explicit definition of the allowed values for a certain datatype
    - detect whether the data conforms to the specific RDF pattern and that the “types” are defined for specific resources,
    - use of different outlier techniques and clustering for detecting wrong values
  - syntactic rules (type of characters allowed and/or the pattern of literal values)
- SV3: detection of ill-typed literals

Semantic accuracy refers to the degree
to which data values correctly represent the real world facts.

Semantic Accuracy metrics

- SA1: detection of outliers
- SA2: detection of inaccurate values
- SA3: detection of inaccurate annotations, labellings or classifications
- SA4: detection of misuse of properties

Consistency means that a knowledge base is free of (logical/formal) contradictions with respect to particular knowledge representation and inference mechanisms.

Consistency metrics

- CS1: detection of use of entities as members of disjoint classes using the formula
- CS2: detection of misplaced classes or properties
- CS3: detection of misuse of owl:DatatypeProperty or owl:ObjectProperty
- CS4: detection of use of members of owl:DeprecatedClass or owl:DeprecatedProperty
- CS5: detection of bogus owl:InverseFunctionalProperty values
- CS6: detection of inconsistencies in spatial data through semantic and geometric constraints
- CS7: the attribution of a resource's property (with a certain value) is only valid if the resource (domain), value (range) or literal value (rdfs ranged) is of a certain type - detected by use of SPARQL queries as a constraint
- CS8: detection of inconsistent values by the generation of a particular set of schema axioms for all properties in a dataset and the manual verification of these axioms

Conciseness refers to the minimization of redundancy of entities at the schema and the data level. Conciseness is classified into

- (i) intensional conciseness (schema level) which refers to the case when the data does not contain redundant schema elements (properties and classes) and
- (ii) extensional conciseness (data level) which refers to the case when the data does not contain redundant objects (instances).

Conciseness metrics

- CN1: intensional conciseness measured by
- (no. of unique properties or classes of a dataset / total no. of properties/classes in a target schema)
- CN2: extensional conciseness measured by:
  - no. of unique instances of a dataset / total number of instances representations in the dataset
  - 1 - (total no. of instances that violate the uniqueness rule / total no. of relevant instances)

Completeness refers to the degree to which all required information is present in a particular dataset. In terms of LD, completeness comprises of the following aspects:

- Schema completeness, the degree to which the classes and properties of an ontology are represented, thus can be called “ontology completeness”,
- Property completeness, measure of the missing values for a specific property,
- Population completeness is the percentage of all real-world objects of a particular type that are represented in the datasets and
- Interlinking completeness, which has to be considered especially in LD, refers to the degree to which instances in the dataset are interlinked.

Completeness metrics

- CM1: schema completeness:
  - no. of classes and properties represented / total no. of classes and properties
- CM2: property completeness:
  - no. of values represented for a specific property / total no. of values for a specific property
- CM3: population completeness:
  - no. of real-world objects represented / total no. of real-world objects
- CM4: interlinking completeness:
  - no. of instances in the dataset that are interlinked
  - percentage of mappable types in a datasets that have not yet been considered in the linksets when assuming an alignment among types

Quality Dimensions

- Accessibility
- Intrinsic
- **Contextual**
- Representational

Contextual Dimensions

- Relevancy
- Trustworthiness
- Understandability
- Timeliness

Relevancy refers to the provision of information which is in accordance with the task at hand and important to the users' query.

Relevancy

- R1: obtaining relevant data by:
  - ranking, which determines the centrality of RDF documents and statements,
  - via crowdsourcing
- R2: measuring the coverage (i.e. number of entities described in a dataset) and level of detail (i.e. number of properties) in a dataset to ensure that the data retrieved is appropriate for the task at hand

Trustworthiness is defined as the degree to which the information is accepted to be correct, true, real and credible.

Trustworthiness metrics

- T1: computing statement trust values based on:
  - provenance information
  - opinion-based method
  - provenance information and trust annotation in Semantic Web-based social-networks
  - annotating triples with provenance data and usage of provenance history to evaluate the trustworthiness of facts
- T2: using annotations for data to encode two blacklists and authority
- ...

Understandability refers to the ease with which data can be comprehended without ambiguity and be used by a human information consumer.

Understandability metrics

- U1: detection of human-readable labelling of classes, properties and entities as well as indication of metadata (e.g. name, description, website) of a dataset
- U2: detect whether the pattern of the URIs is provided
- U3: detect whether a regular expression that matches the URIs is present
- U4: detect whether examples of SPARQL queries are provided
- U5: checking whether a list of vocabularies used in the dataset is provided
- U6: checking the effectiveness and the efficiency of the usage of the mailing list and/or the message boards

Timeliness measures how up-to-date data is relative to a specific task.

Timeliness metrics

- TI1: detecting freshness of datasets based on currency and volatility
- TI2: detecting freshness of datasets based on their data source by measuring the distance between the last modified time of the data source and last modified time of the dataset

Quality Dimensions

- Accessibility
- Intrinsic
- Contextual
- **Representational**

Representational dimensions

- Representational-conciseness
- Interoperability
- Interpretability
- Versatility

Representational conciseness refers to the representation of the data, which is compact and well formatted on the one hand and clear and complete on the other hand.

Representational conciseness

- RC1: detection of long URIs or those that contain query parameters
- RC2: detection of RDF primitives i.e. RDF reification, RDF containers and RDF collections

Interoperability refers to the degree to which the format and structure of the information conforms to previously returned information as well as data from other sources.

Interoperability metrics

- IO1: detection of whether existing terms from all relevant vocabularies for that particular domain have been reused
- IO2: usage of relevant vocabularies for that particular domain

Interpretability refers to technical aspects of the data, that is, whether information is represented using an appropriate notation and whether the machine is able to process the data.

Interpretability metrics

- IN1: identifying objects and terms used to define these objects with globally unique identifiers
- IN2: detecting the use of appropriate language, symbols, units, datatypes and clear definitions
- IN3: detection of invalid usage of undefined classes and properties (i.e. those without any formal definition)
- IN4: detecting the use of blank nodes

Versatility refers to the availability of data in different representations and in an internationalized way.

Versatility metrics

- V1: checking whether data is available in different serialization formats
- V2: checking whether data is available in different languages

Course Feedback