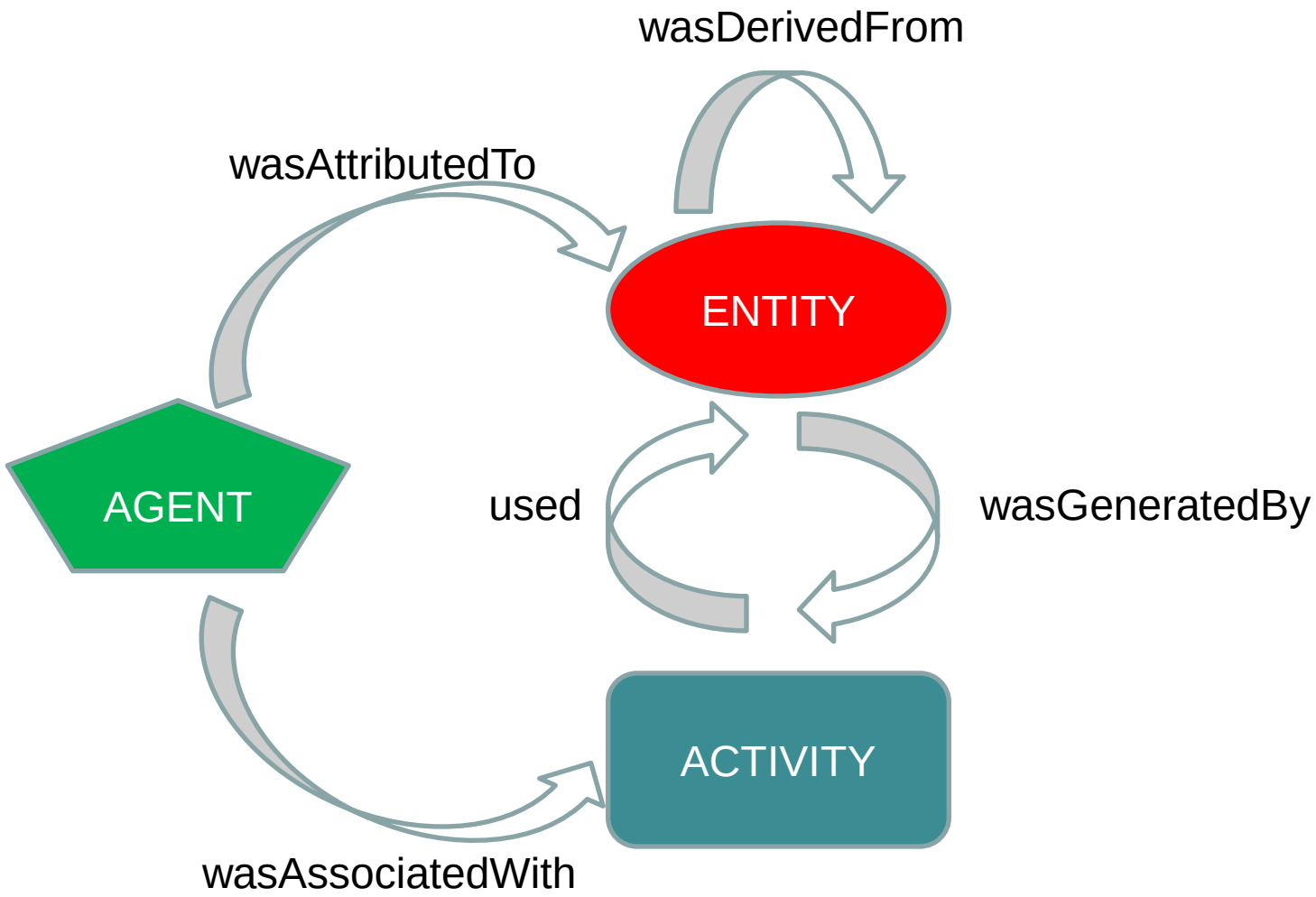
C. Batini & M. Scannapieco
Data and Information Quality Book Figures

Chapter 14 - Quality of Web Data and Quality of Big Data: Open Problems

Comprehensive list of IQ metrics for trust dimensions

Dimension	SubDimension	Description
Believability	computing the trustworthiness of RDF statements	computing a trust value based on user-based ratings or opinion-based method [301]
	computing the trust of an entity	construction of decision networks informed by provenance graphs [257]
	accuracy of computing the trust between two entities	by using a combination of (1) propagation algorithm which utilizes statistical techniques for computing trust values between 2 entities through a path and (2) an aggregation algorithm based on a weighting mechanism for calculating the aggregate value of trust over all paths [570]
	acquiring content trust from users	based on associations that transfer trust from entities to resources [266]
	detection of trustworthiness, reliability and credibility of a data source	use of trust annotations made by several individuals to derive an assessment of the sources' trustworthiness, reliability and credibility [267]
	assigning trust values to data/sources/rules	use of trust ontologies that assign content-based or metadata-based trust values that can be transferred from known to unknown data [338]
	determining trust value for data	using annotations for data such as (i) blacklisting, (ii) authoritativeness and (iii) ranking and using reasoning to incorporate trust values to the data [85]
	meta-information about the identity of information provider	checking whether the provider/contributor is contained in a list of trusted providers [79]
Verifiability	verifying publisher information	stating the author and his contributors, the publisher of the data and its sources [242]
	verifying authenticity of the dataset	whether the dataset uses a provenance vocabulary, eg. the use of the Provenance Vocabulary [242]
	verifying correctness of the dataset	with the help of unbiased trusted third party [79]
	verifying usage of digital signatures	signing a document containing an RDF serialisation or signing an RDF graph [242]
Reputation	reputation of the publisher	survey in a community questioned about other members [266]
	reputation of the dataset	analyzing references or page rank or by assigning a reputation score to the dataset [438]

Key concepts of the PROV Family of Documents



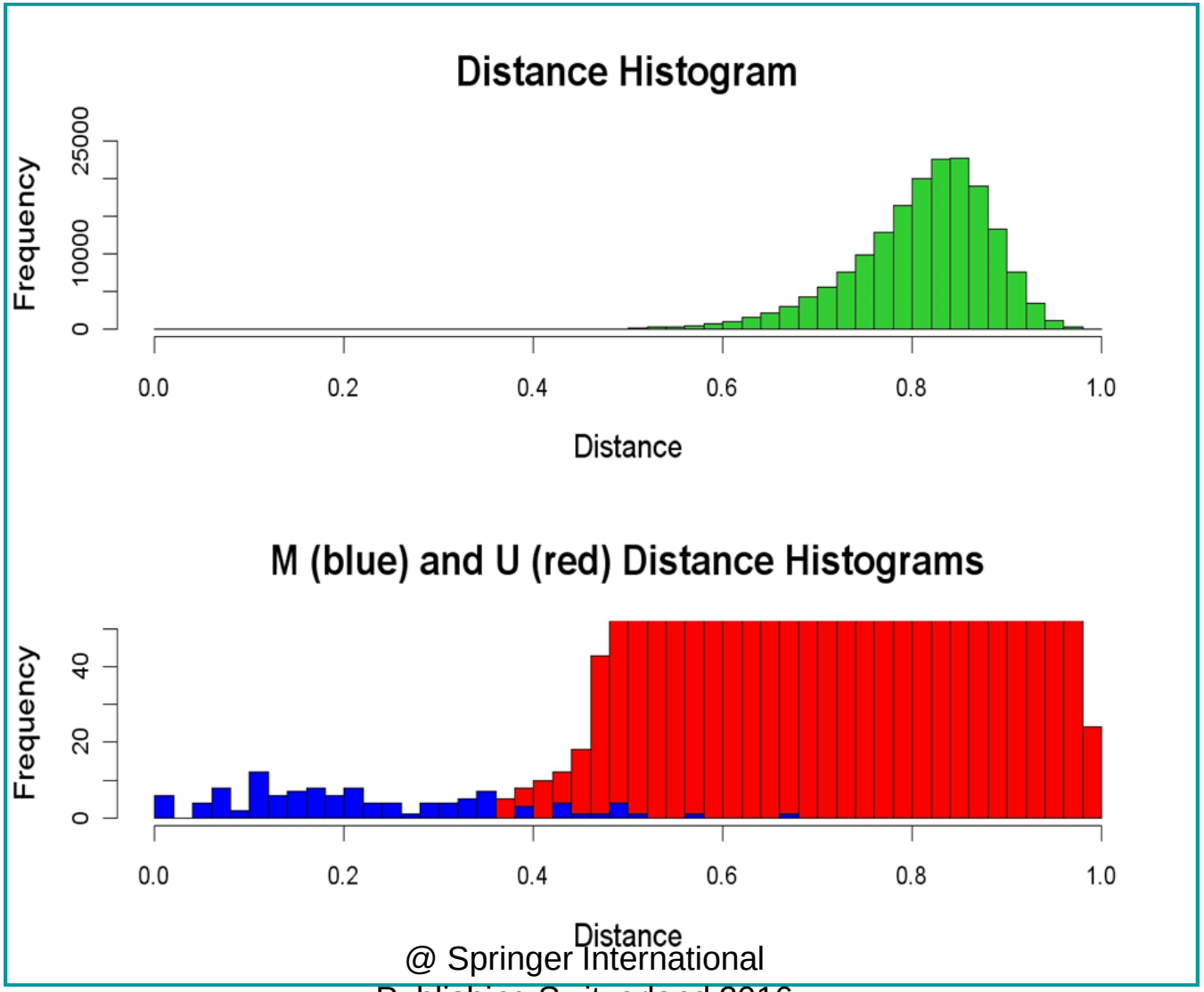
Dimensions of Provenance of Web Data

Category	Dimension	Description
Content	Attribution	Provenance as the sources or entities that were used to create a new result *Responsibility*: Knowing who endorses a particular piece of information or result *Origin*: recorded vs. reconstructed, verified vs. non-verified, asserted vs. inferred
	Process	Provenance as the process that yielded an artifact *Reproducibility* (e.g. workflows, mashups, text extraction) *Data Access* (e.g. access time, accessed server, party responsible for accessed server)
	Evolution and versioning	*Republishing* (e.g. re-tweeting, re-blogging) *Updates* (e.g. a document with content from various sources and that changes over time)
	Justification for decisions	Includes argumentation, hypotheses, why-not questions
	Entailment	Given the results to a particular query, what axioms or tuples led to those result
Management	Publication	Making provenance information available (expose, distribute)
	Access	Finding and querying provenance information
	Dissemination Control	Track policies specified by creator for when/how an artifact can be used *Access Control*: incorporate access control policies to access provenance information *Licensing*: stating what rights the object creators and users have based on provenance *Law enforcement*(e.g. enforcing privacy policies on the use of personal information)
	Scale	how to operate with large amounts of provenance information
Use	Understanding	End user consumption of provenance *Abstraction*: multiple levels of description, summary *Presentation, Visualization*
	Interoperability	Combining provenance produced by multiple different systems
	Comparison	Finding what is in common in the provenance of two or more entities (e.g. two experimental results)
	Accountability	The ability to check the provenance of an object with respect to some expectation *Verification* of a set of requirements *Compliance* with a set of policies
	Trust	Making trust judgments based on provenance *Information quality* *Reputation, Reliability*
	Imperfections	Reasoning about provenance information that is not complete or correct *Incomplete provenance* *Uncertain,probabilistic provenance* *Erroneous provenance* *Fraudulent provenance*
	Debugging	Using provenance to detect bugs or failures of processes.

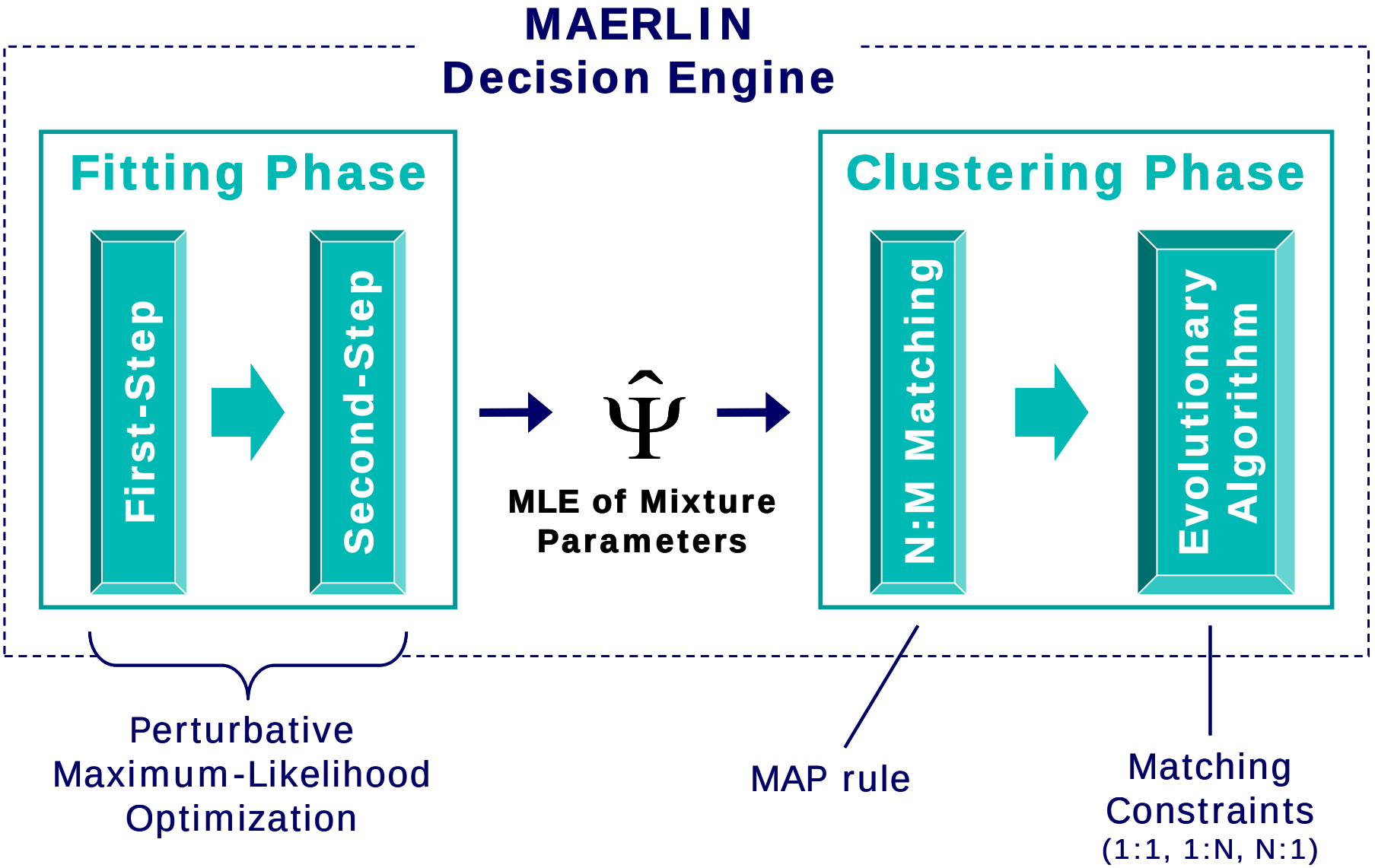
PROV-O document that defines provenance information for Linked open data

Audience	Document Name	Description
Users	Prov-Primer	It is the entry point to PROV offering an introduction to the provenance data model. This is where you should start and for many may be the only document needed.
Developers	Prov-O	It defines a light-weight OWL2 ontology for the provenance data model. This is intended for the Linked Data and Semantic Web community.
	Prov-XML	Defines an XML schema for the provenance data model. This is intended for developers who need a native XML serialization of the PROV data model.
	Prov-AQ	Defines how to use Web-based mechanisms to locate and retrieve provenance information.
	Prov-DC	Defines a mapping between Dublin Core and PROV-O
	Prov-Dictionary	Defines constructs for expressing the provenance of dictionary style data structures.
Advanced	Prov-DM	It defines a conceptual data model for provenance including UML diagrams. PROV-O, PROV-XML and PROV-N are serializations of this conceptual model.
	Prov-N	Defines a human-readable notation for the provenance model. This is used to provide examples within the conceptual model as well as used in the definition of PROV-CONSTRAINTS.
	Prov-CONSTRAINTS	Defines a set of constraints on the PROV data model that specifies a notion of valid provenance. It is specifically aimed at the implementors of validators.
	Prov-Sem	Defines a declarative specification in terms of first-order logic of the PROV data model.
	Prov-LINKS	Defines extensions to PROV to enable linking provenance information across bundles of provenance descriptions.

Distance histogram and matches and non-matches histograms: the nonmatches red histogram in the lower panel has been cut to allow the detection of the very small matches distribution (blue)



MAERLIN decision engine



Various types of consistency as defined in [562]

Types of Consistency	Numerical/ Temporal/ Frequency	Individual Data/ Data Streams/ Both	Definition
Numerical	Numerical	Individual Data	Collected data should be accurate
Temporal	Temporal	Individual Data	Data should be delivered to the sink before or by it is expected
Frequency	Frequency	Both	Controls the frequency of dramatic data changes and abnormal readings of data streams
Absolute numerical	Numerical	Both	Sensor reading is out of the normal range, which can be preset by the application
Relative numerical	Numerical	Both	Error between the real field reading and the corresponding data at the sink
Hop	Numerical	Individual Data	Data should keep consistency at each hop
Single path	Numerical and Temporal	Individual Data	Consistency holds when data are transmitted from the source to the sink using a single path
Multiple path	Numerical and Temporal	Individual Data	Consistency holds when data are transmitted from the source to the sink using multiple paths
Strict	Numerical and Temporal	Data Streams	Differs from hope consistency because it is defined on a set of data and requires no data loss
Alpha-loss	Numerical and Temporal	Data Streams	Similar to strict consistency except that alfa-data loss are accepted at the sink
Partial	Numerical and Temporal	Data Streams	Similar to alfa consistency except that temporal consistency is released
Trend	Numerical and Temporal	Data Streams	Similar to partial consistency except that numerical consistency is released
Range frequency	Frequency	Data Streams	The number of abnormal readings exceed a certain number preset by the application
Change frequency	Frequency	Data Streams	Changes of sensor readings exceeds preset threshold

Clusters, quality of context dimensions, definitions in [427] and related sources of context data

Cluster	Dimension in Cluster	Definition	Sources of QoC used in the evaluation
Accuracy	Up-to-Dateness	Degree of rationalism to us e acontext object for a specific application at a given time	Measurement Time Current Time
Accuracy	Precision	–	–
Completeness	Completeness	Quantity of information that is provided for a specific object	Ratio of number of attributes filled to the total number of attributes
Completeness	Significance	Worth or preciousness of the context information in a specific situation	Critical value
Redundancy	Conciseness	–	–
Consistency	Representation Consistency	–	–
Trustworthiness	Trustworthiness	Belief that we have in the correct information in a given context object	Source location Information entity location Sensor data accuracy

Results of the application of Naive Bayes to the complete set of questions related to Web sales

Question	Precision	Sensitivity	Specificity	Proportion Web sales = Yes (observed)	Proportion Web sales = Yes (predicted)
Web sales functionality	0.78	0.50	0.86	0.21	0.21
Orders tracking	0.82	0.49	0.85	0.18	0.11
Description and price list of goods	0.62	0.44	0.79	0.48	0.32
Personalised content for regular visitors	0.74	0.41	0.781	0.09	0.23
Possibility to customise online goods	0.86	0.53	0.87	0.05	0.14
Privacy policy statement	0.59	0.57	0.64	0.68	0.51
Online job application	0.69	0.521	0.78	0.35	0.33