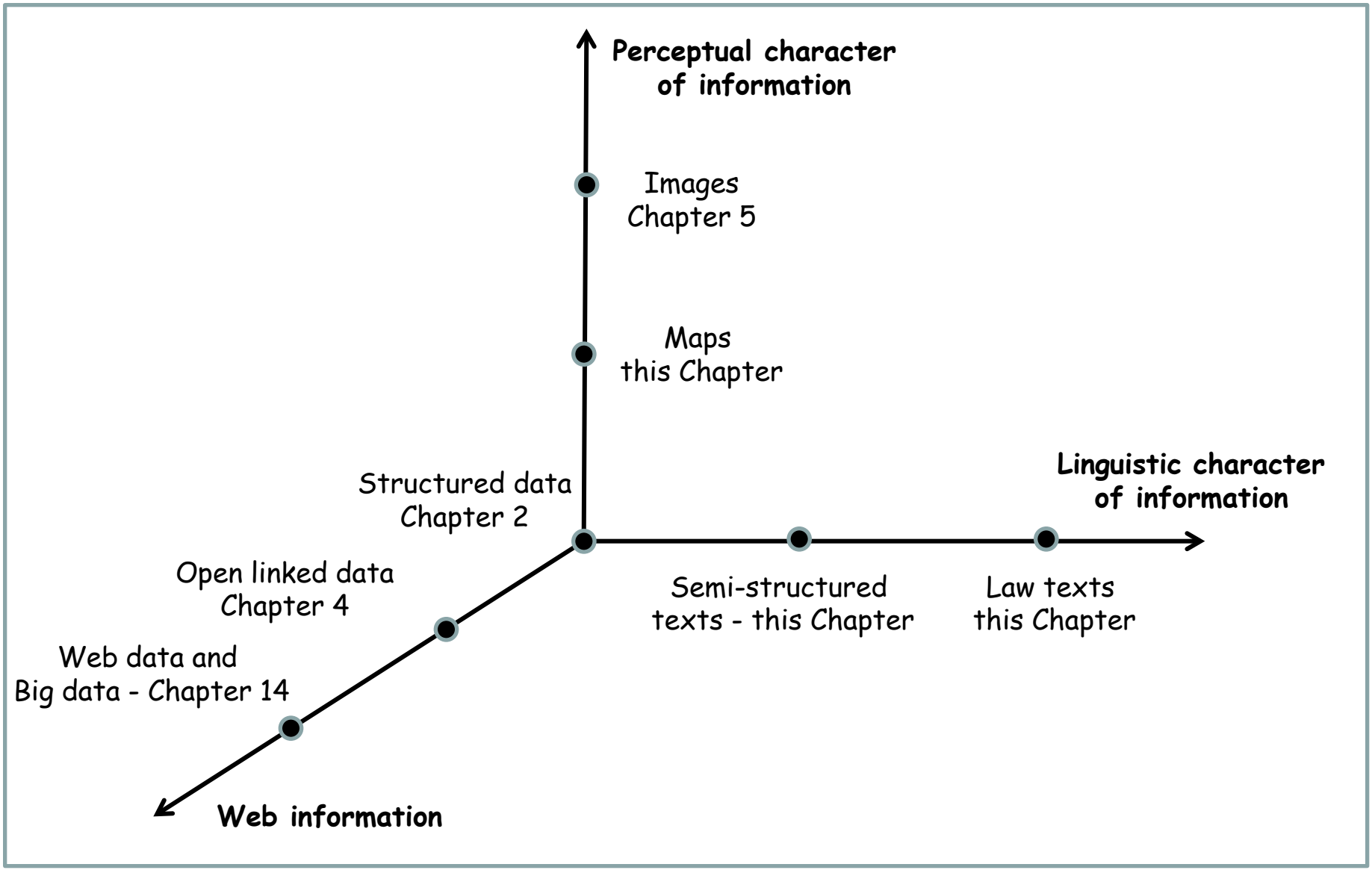
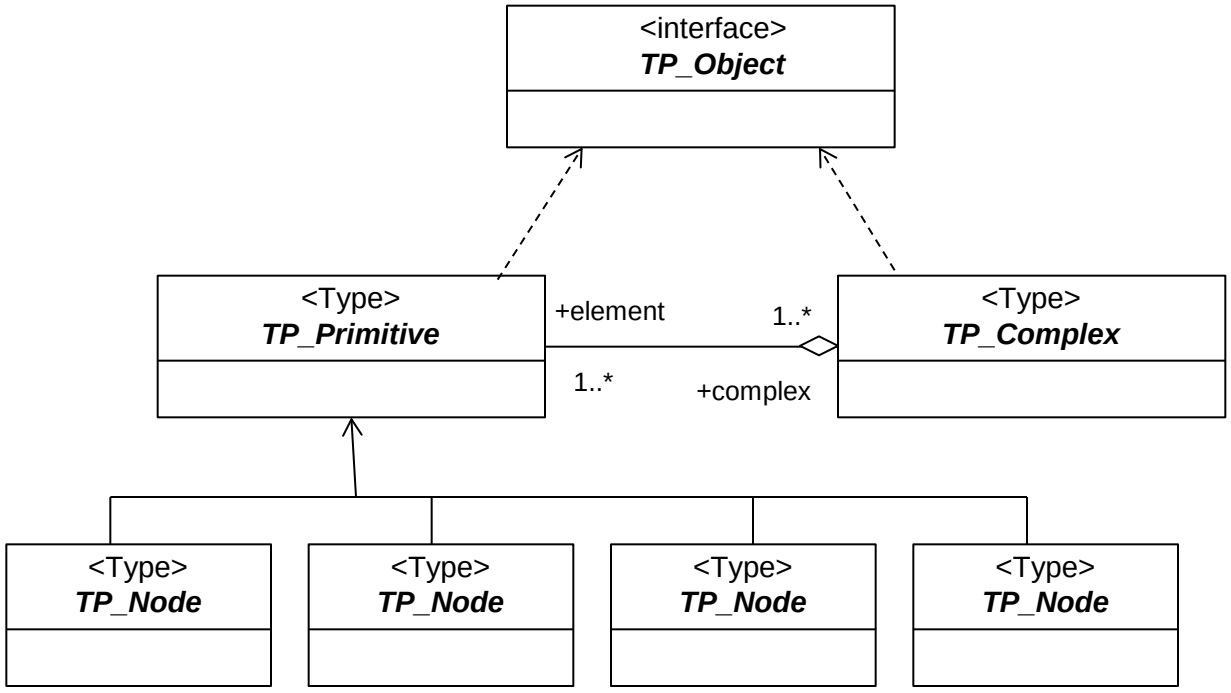
C. Batini & M. Scannapieco
Data and Information Quality Book Figures

Chapter 3: Information Quality Dimensions for Maps and Texts

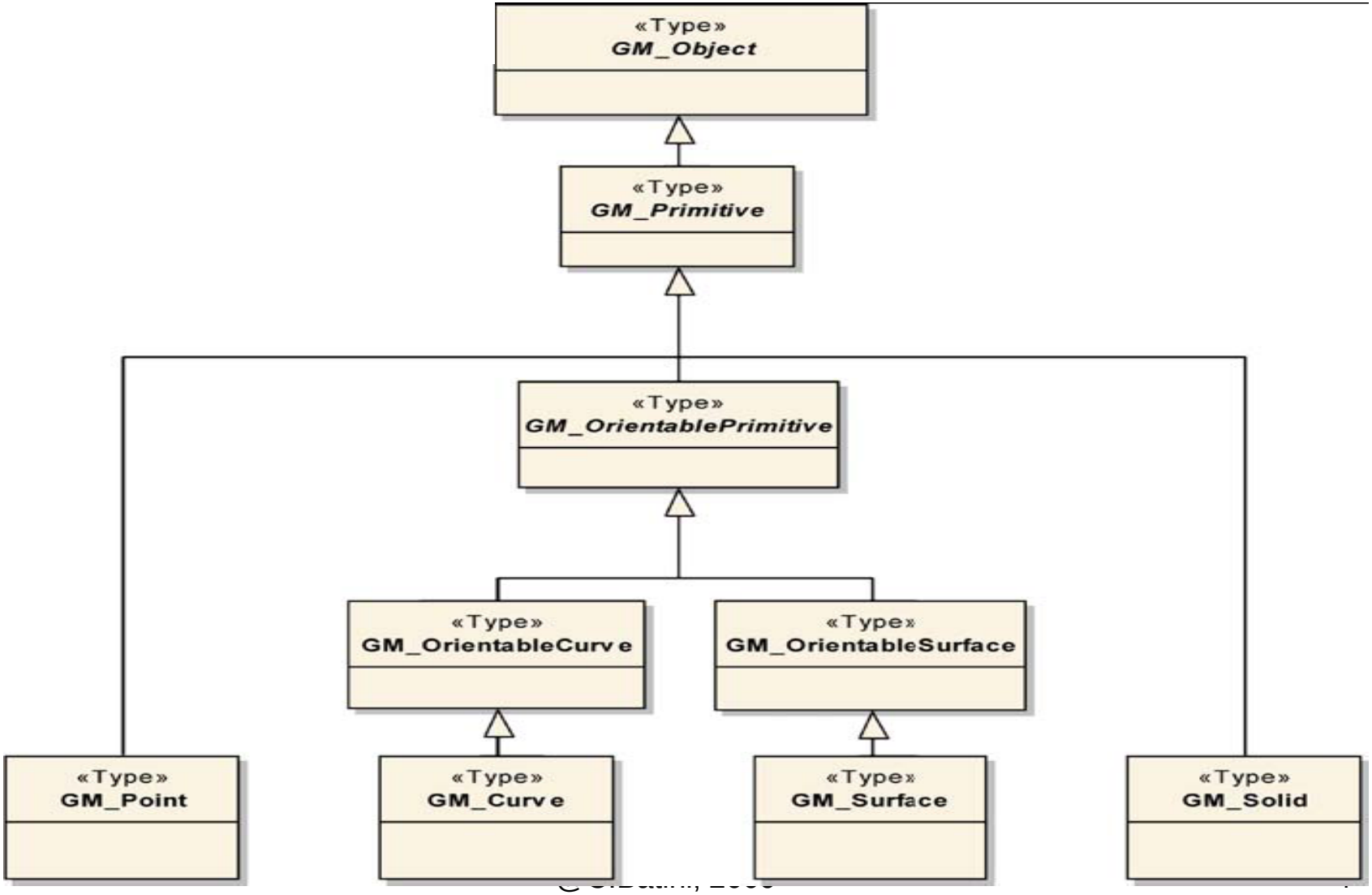
Types of information considered in the book according to the perceptual, linguistic, and Web coordinates



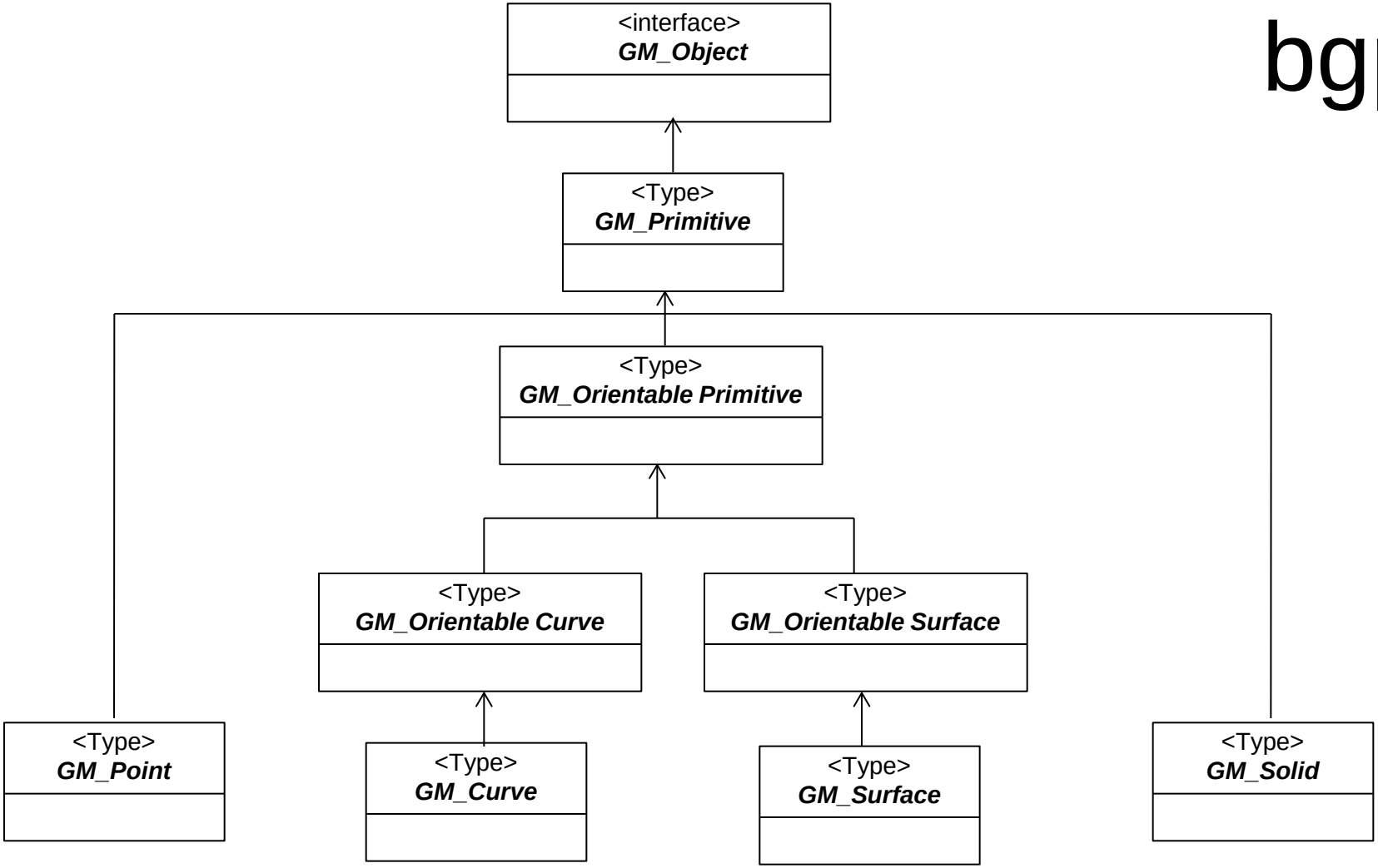
Schema of topological primitives specified by ISO 19107



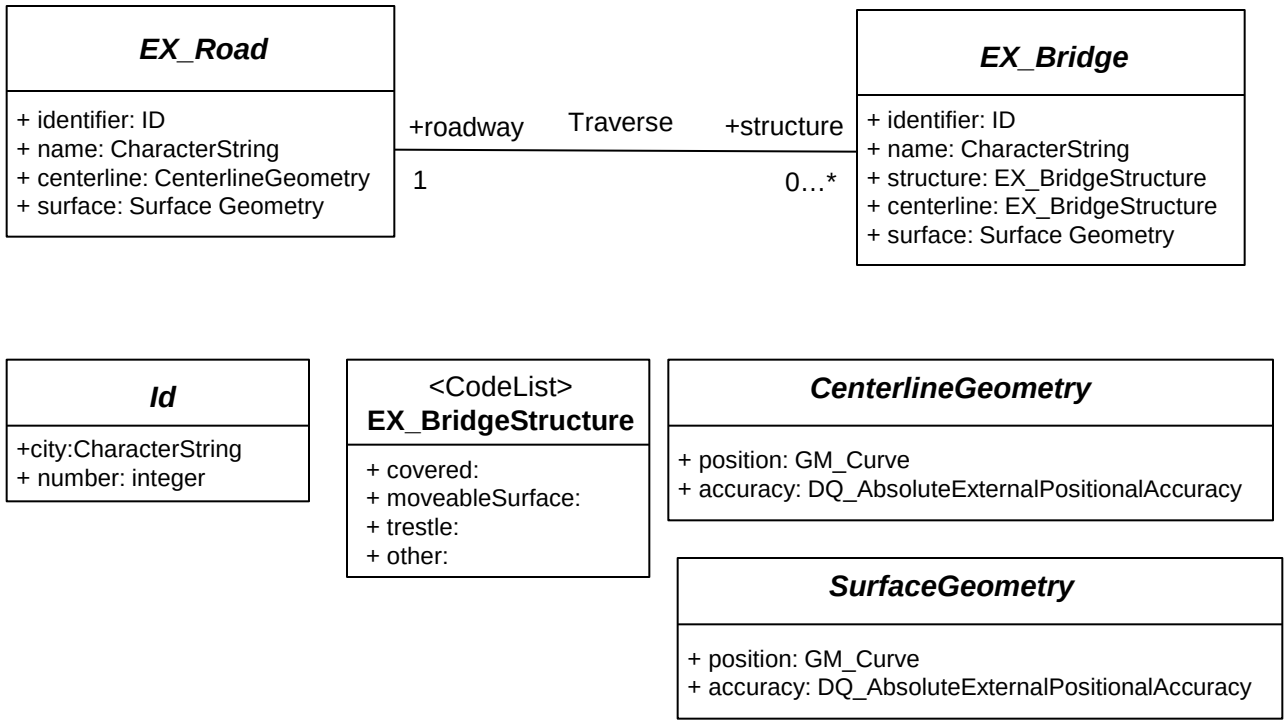
Schema of basic geometric primitives specified by ISO 19107



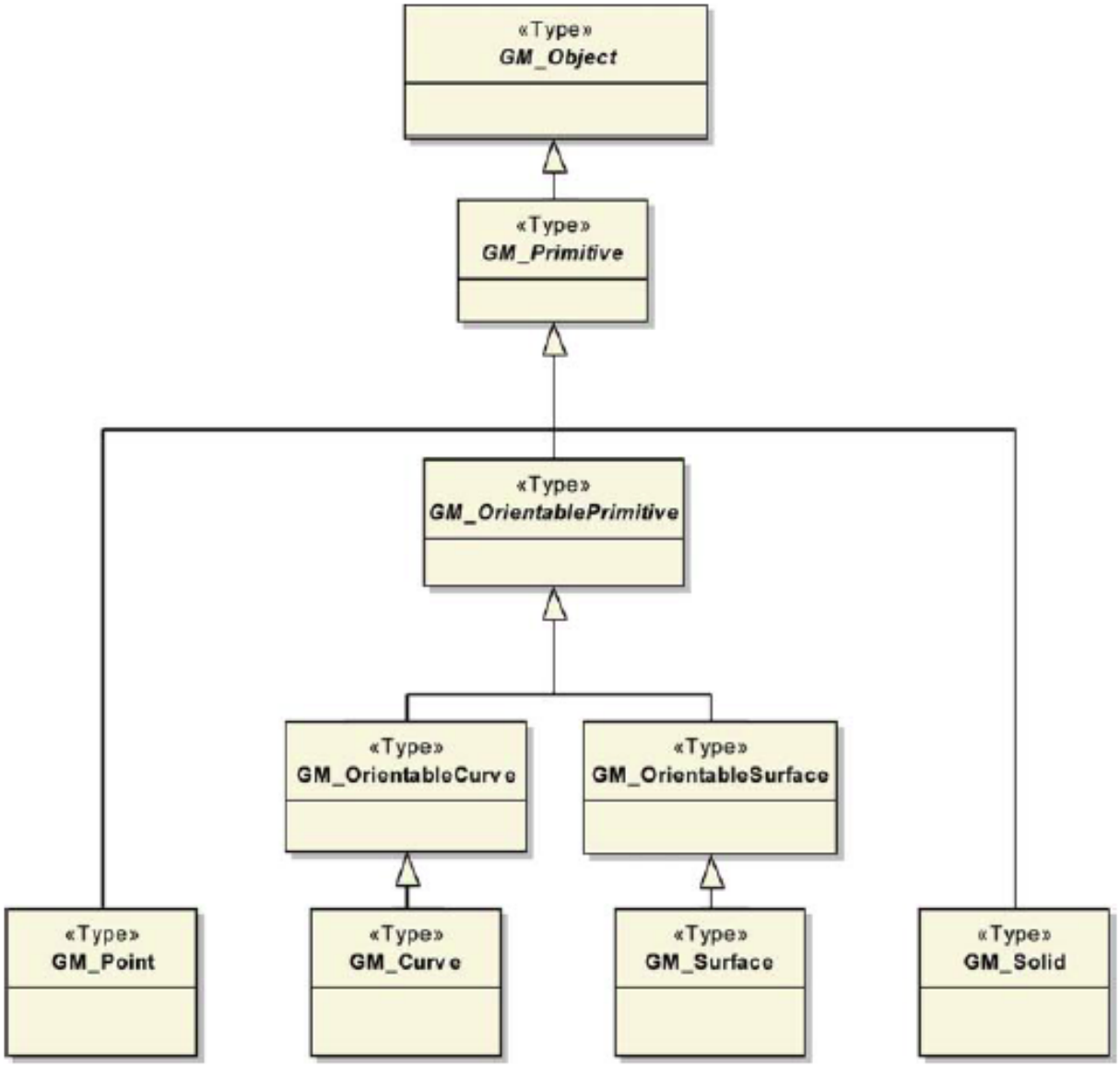
bgp


<interface>
GM_Object
<Type>
GM_Primitive
<Type>
GM_Orientable Primitive
<Type>
GM_Orientable Curve
<Type>
GM_Orientable Surface
<Type>
GM_Point
<Type>
GM_Curve
<Type>
GM_Surface
<Type>
GM_Solid

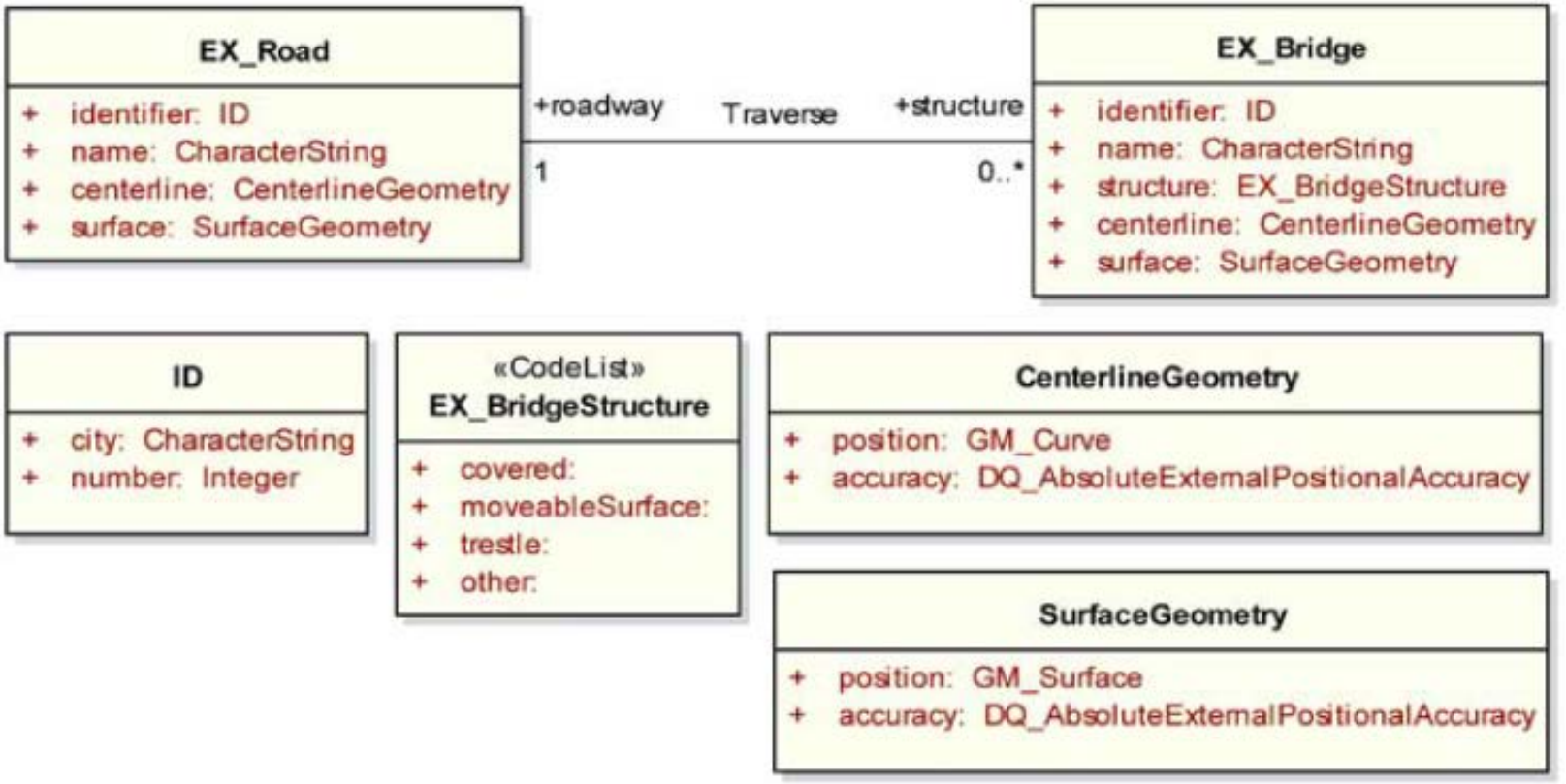
Application schema for representing roads and bridges compliant with ISO 19109 rules



bgp


«Type»
GM_Object
«Type»
GM_Primitive
«Type»
GM_OrientablePrimitive
«Type»
GM_OrientableCurve
«Type»
GM_OrientableSurface
«Type»
GM_Point
«Type»
GM_Curve
«Type»
GM_Surface
«Type»
GM_Solid

Modeling with ISO 191xx Standards
Road-bridge application schema example



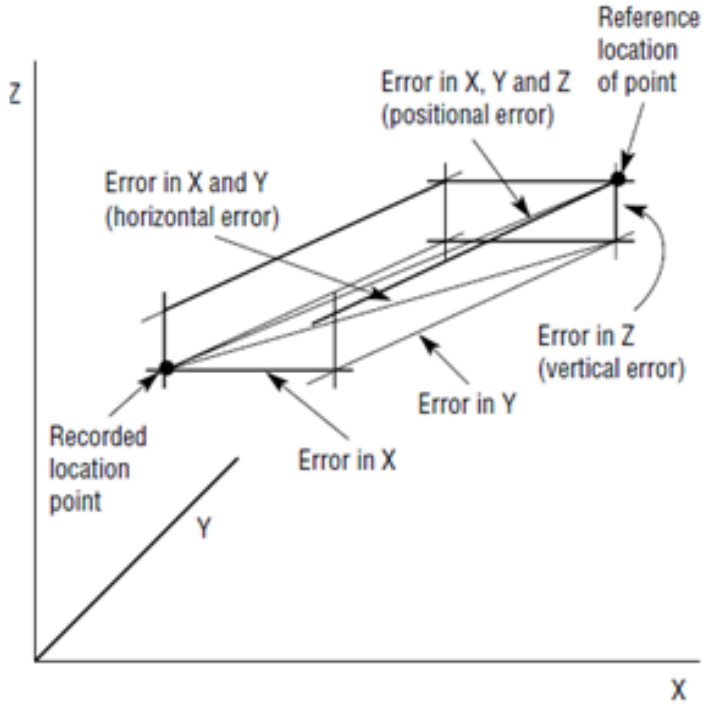
Quality dimensions of maps in the ISO 19100 geographic information quality standards and in the spatial data transfer standard

Cluster	Dimension	Source	Definition
Accuracy	Positional	Iso 19100	Accuracy of the position of features
Accuracy	Relative positional	Iso 19100	Closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true
Accuracy	Absolute positional	Iso 19100	Closeness of reported coordinate values to values accepted as or being true
Accuracy	Horizontal positional	SDTS	Accuracy of the horizontal position in the data set
Accuracy	Vertical positional	SDTS	Accuracy of the vertical position in the data set
Accuracy	Gridded data position	Iso 19100	Closeness of gridded data position values to values accepted as or being true
Accuracy	Thematic	Iso 19100	Accuracy of quantitative attributes and the correctness of non quantitati-ve attributes and of the classifications of features and their relationships
Accuracy	of quantitative attributes	Iso 19100	Accuracy of quantitative attributes
Accuracy	Temporal validity	Iso 19100	Validity of data with respect to time
Accuracy	of a time measurement	Iso 19100	Correctness of the temporal references of an item (reporting of error in time measurement)
Accuracy	Correctness of non quantitative attributes	Iso 19100	Correctness of non-quantitative attributes
Correctness	Correctness of classification	Iso 19100	Comparison of the classes assigned to features or their attributes to a universe of discourse (e.g. ground truth or reference data set)
Completeness	-	Iso 19100	Presence or absence of features, heir attributes and relationships
Completeness	Pertinence (or Commission)	Iso 19100	Excess data present in a dataset
Consistency	Logical	Iso 19100	Degree of adherence to logical rules of data structure, attribution and relationships
Consistency	Conceptual	Iso 19100	Adherence to rules of the application conceptual schema
Consistency	Domain	Iso 19100	Adherence of values to the value domains

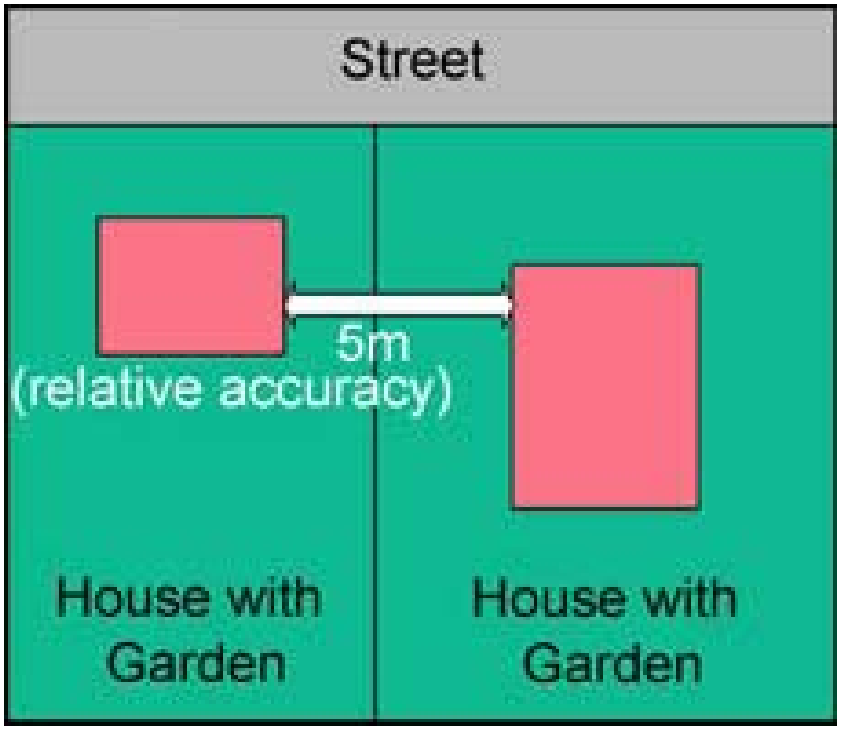
dimvsmapscharacteristics

Conceptual issue → Dimension Cluster ↓	Space -topological	Space - geometric	Space -thematic	Temporal
Accuracy		1. Positional 2. Absolute position acc. 3. Relative position acc. 4. Gridded data pos.acc. 5. Horizontal acc. 6. Vertical acc. 7. Geometric precision	1. Thematic acc. 2. Accuracy/corr. of quantitative attributes 3. Accuracy/corr. of non quantitative attributes 4. Classification accuracy/correctness 5. Thematic precision	1. of a time measurement 2. Temporal validity 3. Temporal precision
Completeness			1. Completeness 2. Pertinence	
Consistency	1. Conceptual 2. Topological	Conceptual	1. Logical 2. Conceptual 3. Domain 4. Format	Temporal

geomaccs

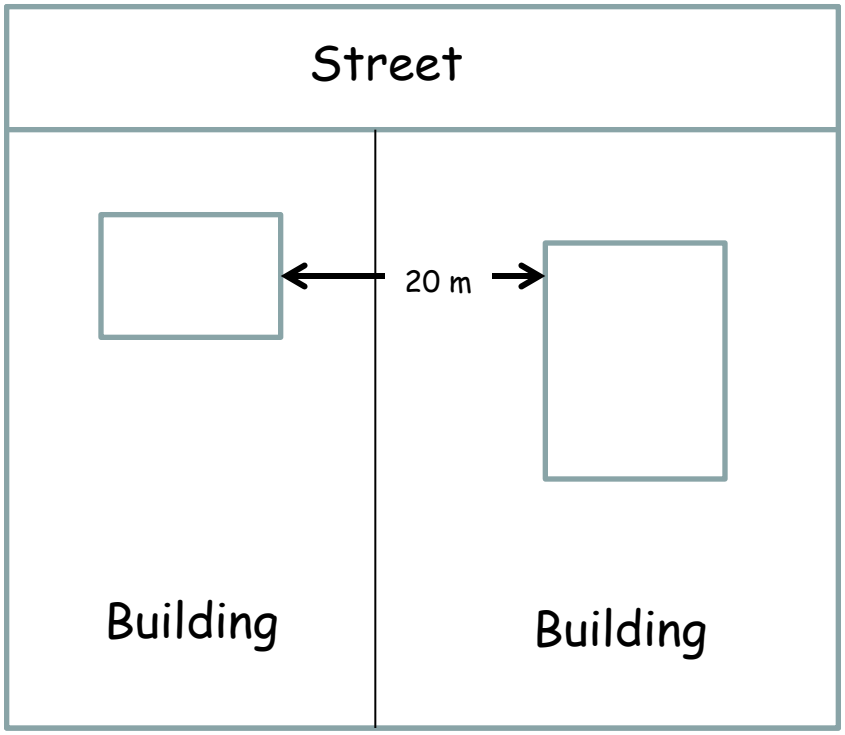

Reference location of point
Error in X, Y and Z (positional error)
Error in X and Y (horizontal error)
Error in Z (vertical error)
Error in Y
Error in X
Recorded location point
Z
Y
X

Relative positional accuracy

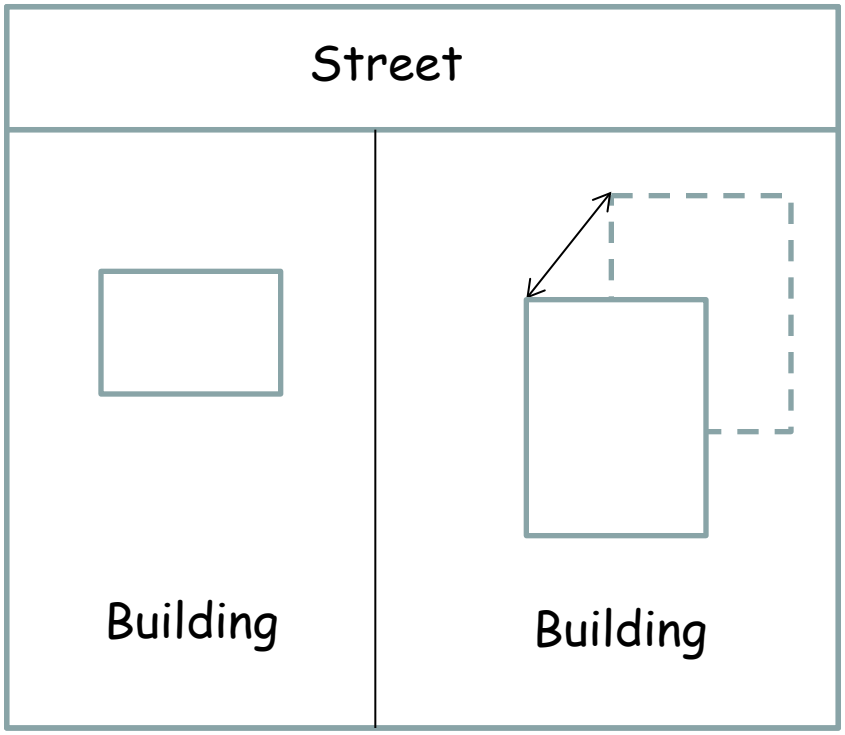


Relative Positional Accuracy, which has traditionally been used to indicate the positional accuracy of maps, is defined as the difference of the distance between two defined points in a geospatial dataset and the true distance between these points within the overall reference system.

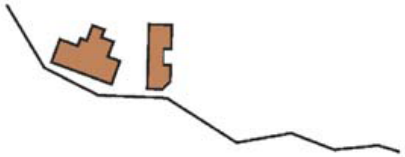
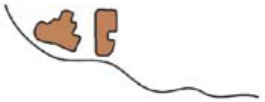
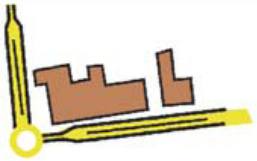
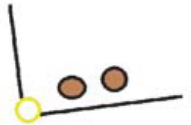
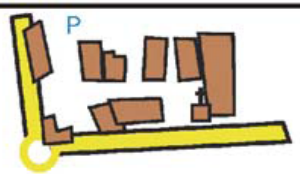
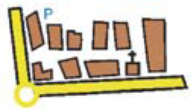
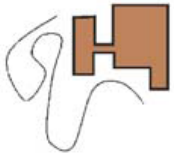
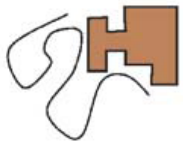
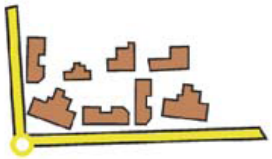
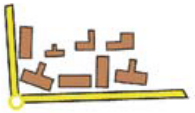
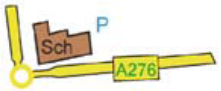
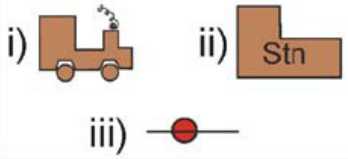
exampleofrelposaccuracy



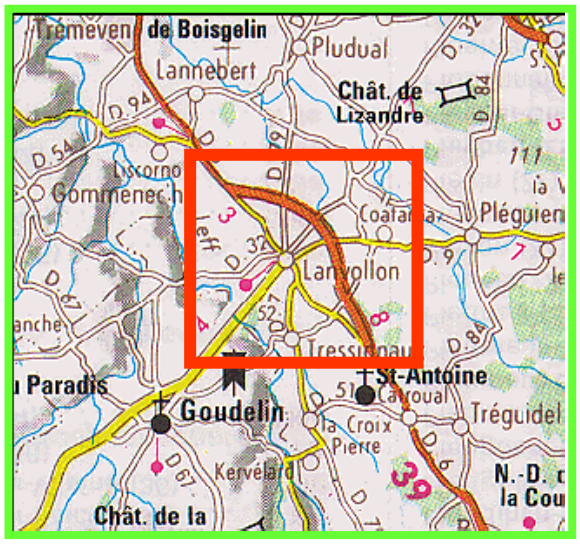
exampleofabsposaccuracy

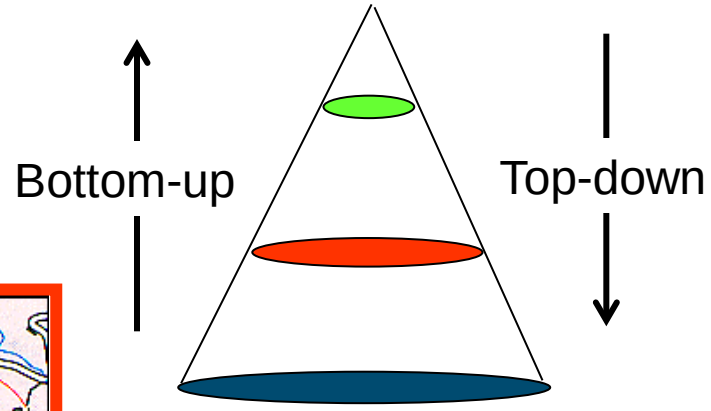


Generalization and Symbolization
Cartographic generalization operations

Operator	Before	After
(a) **Smoothing** Reduce angularity of the map object.		
(b) **Collapse** Reduce dimensionality of map object (area to point, linear polygon to line).		
(c) **Displacement** Small movement of map objects in order to minimise overlap.		
(d) **Enhancement** Emphasize characteristics of map feature and meet minimum legibility requirements.		
(e) Typification Replacement of a group of map features with a prototypical subset.		
(f) Text Placement Non overlapping unambiguous placement of text.		
(g) Symbolization Change of symbology according to theme (pictorial, iconic), or reduce space required for symbol.		

absqvdmapsq





Quality dimensions of loosely structured text

A fully unstructured (??) text

I saw them not long ago I love flowers Id love to have the whole place swimming in roses God of heaven theres nothing like nature the wild mountains then the sea and the waves rushing then the beautiful country with the fields of oats and wheat and all kinds of things and all the fine cattle going about that would do your heart good to see rivers and lakes and flowers all sorts of shapes and smells and colours springing up even out of the ditches primroses and violets nature it is as for them saying theres no God I wouldnt give a snap of my two fingers for all their learning why dont they go and create something I often asked him atheists or whatever they call themselves go and wash the cobbles off themselves first then they go howling for the priest and they dying and why why because theyre afraid of hell on account of their bad conscience ah yes I know them well who was the first person in the universe before there was anybody that made it all who ah that they dont know neither do I so there you are they might as well try to stop the sun from rising tomorrow the sun shines for you he said the day we were lying among the rhododendrons on Howth head in the grey tweed suit and his straw hat the day I got him to propose to me yes first I gave him the bit of seedcake out of my mouth and it was leapyear like now yes 16 years ago my God after that long kiss I near lost my breath yes he said I was a flower of the mountain yes so we are flowers all a womans body yes that was one true thing he said in his life and the sun shines for you today yes that was why I liked him because I saw he understood or felt what a woman is and I knew I could always get round him and I gave him all the pleasure I could leading him on till he asked me to say yes and I wouldnt answer first only looked out over the sea and the sky I was thinking of so many things he didnt know of Mulvey and Mr Stanhope and Hester and father and old captain Groves and the sailors playing all birds fly and I say stoop and washing up dishes they called it on the pier and the sentry in front of the governors house with the thing round his white helmet poor devil half roasted and the Spanish girls laughing in their shawls and their tall combs and the auctions in the morning the Greeks and the jews and the Arabs and the devil knows who else from all the ends of Europe and Duke street and the fowl market all clucking outside Larby Sharons and the poor donkeys slipping half asleep and the vague fellows in the cloaks asleep in the shade on the steps and the big wheels of the carts of the bulls and the old castle thousands of years old yes and those handsome Moors all in white and turbans like kings asking you to sit down in their little bit of a shop and Ronda with the old windows of the posadas 2 glancing eyes a lattice hid for her lover to kiss the iron and the wineshops half open at night and the castanets and the night we missed the boat at Algeciras the watchman going about serene with his lamp and O that awful deepdown torrent O and the sea the sea crimson sometimes like fire and the glorious sunsets and the figtrees in the Alameda gardens yes and all the queer little streets and the pink and blue and yellow houses and the rosegardens and the jessamine and geraniums and cactuses and Gibraltar as a girl where I was a Flower of the mountain yes when I put the rose in my hair like the Andalusian girls used or shall I wear a red yes and how he kissed me under the Moorish wall and I thought well as well him as another and then I asked him with my eyes to ask again yes and then he asked me would I yes to say yes my mountain flower and first I put my arms around him yes and drew him down to me so he could feel my breasts all perfume yes and his heart was going like mad and yes I said yes I will Yes.

- Trieste-Zurich-Paris 1914-1921

A fully unstructured (??) text

lstd

Conceptual issue → Cluster ↓	Lexicon	Syntax	Semantics	Rhetoric	Pragmatics
Accuracy	Lexical accuracy	Syntactic accuracy			
Readability	Readability				
	Text comprehension Closer-to-text base comprehension Closer-to-situation model level comprehension				
Consistency	Coherence Referential Cohesion – local co-reference Referential Cohesion – global co-reference				
Accessibility					Cultural accessibility

fla

$$\text{Lexical accuracy} = \frac{\sum_{i}^{K} \text{closeness}\,(w_i, V)}{K}$$

gunningfox

$$0.4 * \left[\left(\frac{words}{sentence} \right) + 100 * \left(\frac{complexwords}{words} \right) \right]$$

exgfi The Gunning-Fox Index
An example

In **describing** the humpback whale song, we will adhere to the **following designations** . The shortest sound that is **continuous** to our ears when heard in "real time" will be called a "unit." Some units when listened to at slower speeds, or **analyzed** by machine, turn out to be a series of pulses or **rapidly** sequenced, discrete tones. In such cases, we will call each discrete pulse or tone a "subunit." A series of units is called a "phrase." An **unbroken** sequence of **similar** phrases is a "theme," and **several** distinct themes combine to form a "song."

{From "Songs of Humpback Whales." 1971. Payne, R. S. & S. McVay. Science 173: 585-597.}

This passage has seven sentences and 96 words.
The average sentence length (ASL) is 13.7.
There are nine difficult words (in **boldface**).
Gunning's Fox index = 0.4 * (13.7 + 9.375) = 9.23.

ariindex

$$ARI = 4{,}71 * \frac{characters}{words} + 0{,}5 * \frac{complex\ words}{sentences} - 21{,}43$$

Crm Readability metrics: comparison

Index	Interpretation	Peculiarity
Gunning Fox Index	estimate of the grade level required to understand the document (years of education)	It considers complex words
Flesch Kinkaid Grade Level	estimate of the grade level required to understand the document (years of education)	It uses syllables per word
ARI	The same	It uses characters per word
SMOG	The same	It considers complex words
Flesch Reading Ease	General measure of readability	It uses syllables per word.

agatha

Characters in "The Mysterious Affair at Styles"
Captain Hastings, the narrator, on sick leave from the Western Front.
Hercule Poirot, a famous Belgian detective exiled in England; Hastings' old friend
Chief Inspector Japp of Scotland Yard
Emily Inglethorp, mistress of Styles, a wealthy old woman
Alfred Inglethorp, her much younger new husband
John Cavendish, her elder stepson
Mary Cavendish, John's wife
Lawrence Cavendish, John's younger brother
Evelyn Howard, Mrs. Inglethorp's companion
Cynthia Murdoch, the beautiful, orphaned daughter of a friend of the family
Dr. Bauerstein, a suspicious toxicologist

Cognitively motivated features – basic
- Number of words
- Number of sentences
- Number of paragraphs
- Average number of words per sentence
- Average number of sentences per paragraph
- Average number of syllabes per word

Cognitively motivated features – complex
- Incidence of functional words
- Average number or verbs hyperonyms
- Number of person pronouns
- Number of negations
- Number of connectives
- Verb ambiguity ratio
- Nouns ambiguity ratio

Syntactic constructions considered in the text wimplification system
- Incidence of caluses
- Incidence of subordination

Features derived form n-gram language models (LM) plus out-of-vocabulary rate scores
- LM probability of unigrams
- LM probability of bibrams
- LM probaility of trigrams
- Out-of-vocabulary words

- Cognitively motivated features – basic
  - Number of words
  - Number of sentences
  - Number of paragraphs
  - Average number of words per sentence
  - Average number of sentences per paragraph
  - Average number of syllabes per word
- Cognitively motivated features – complex
  - Incidence of functional words
  - Average number or verbs hyperonyms
  - Number of person pronouns
  - Number of negations
  - Number of connectives
  - Verb ambiguity ratio
  - Nouns ambiguity ratio
- Syntactic constructions considered in the text wimplification system
  - Incidence of caluses
  - Incidence of subordination
- Features derived form n-gram language models (LM) plus out-of-vocabulary rate scores
  - LM probability of unigrams
  - LM probability of bibrams
  - LM probability of trigrams
  - Out-of-vocabulary words

The four basic elements of reading ease in \cite{dubay2004principles}

Content Proposition Organization Coherence	**Style** Syntactic and Semantic elements
Structure Chapters Headings Navigation	**Design** Typography Format Illustrations

If

Surface code
- Word composition (graphemes, phonemes, syllables, morphemes, lemmas, tense, aspect)
- Words (lexical items)
- Part of speech categories (noun, verb, adjective, adverb, determiner, connective)
- Syntactic composition (noun-phrase, verb-phrase, prepositional phrases, clause)
- Linguistic style and dialect

Textbase
- Explicit propositions
- Referents linked to referring expressions
- Connectives that explicitly link clauses
- Constituents in the discourse focus versus linguistic presuppositions

Situation model
- Agents, objects, and abstract entities
- Dimensions of temporality, spatiality, causality, intentionality
- Inferences that bridge and elaborate ideas
- Given versus new information
- Images and mental simulations of events
- Mental models of the situation

Genre and rhetorical structure
- Discourse category (narrative, persuasive, expository, descriptive)
- Rhetorical composition (plot structure, claim + evidence, problem + solution, etc.)
- Epistemological status of propositions and clauses (claim, evidence, warrant, hypothesis)
- Speech act categories (assertion, question, command, promise, indirect request, greeting, expressive evaluation)
- Theme, moral, or point of discourse

Pragmatic communication
- Goals of speaker / writer and listener / reader
- Attitudes (humor, sarcasm, eulogy, deprecation)
- Requests for clarification and backchannel feedback (spoken only)

mca

$$CA = \frac{\text{Number of occurences of difficult words}}{\text{Total number of word occurences}}$$

localcoreference Cohesion metrics: referential cohesion: local co-reference

- Co-reference occurs when a noun, pronoun or a noun phrase refers to another constituent in the text.
- A simple measure of co-referential text cohesion is the proportion of adjacent sentence pairs in the text that share a common noun argument:

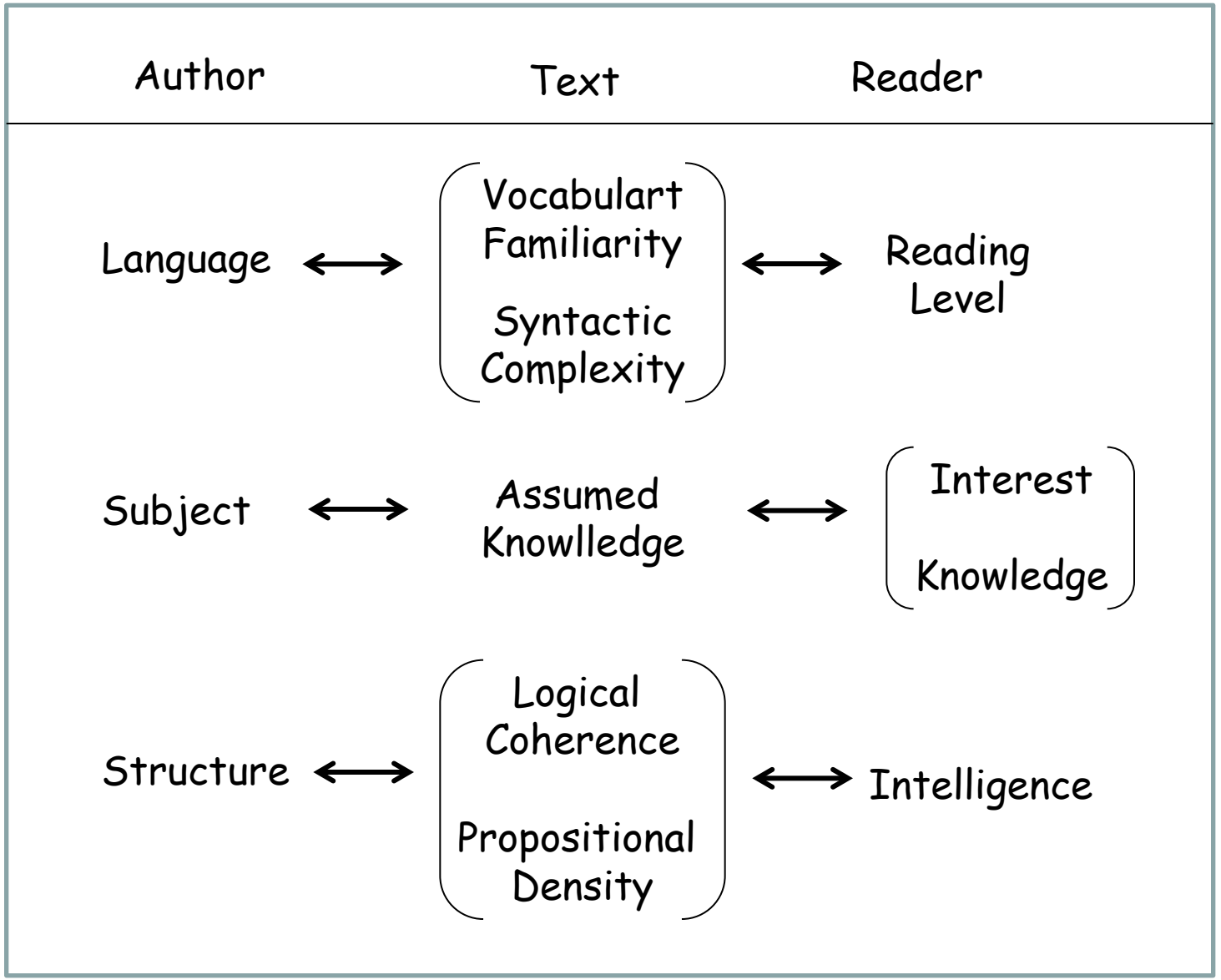
$$\text{Co-reference cohesion local} = \frac{\sum_{i=1}^{n-1} R_{i,i+1}}{n-1}$$

globalcoreference Cohesion metrics: referential cohesion: global co-reference

This measure includes all possible pairs of sentences when co-referential cohesion is computed. The metric is the proportion of pairs that have a co-referential connection:

Co-reference cohesion global = $$: \frac{\sum_{i=1}^{n}\sum_{j=i}^{n} R_{ij} \,|\, i<j}{n \times \frac{n-1}{2}}.$$

Matches needed for easy reading - mer



localcoreference

Co-reference cohesion local =

$$\frac{\sum_{i=1}^{n-1} R_{i,i+1}}{n-1}$$

globalcoreference

Co-reference cohesion global =

$$\frac{\sum_{i=1}^{n}\sum_{j=i}^{n} R_{ij} \mid i<j}{n \times \frac{n-1}{2}}.$$

Quality of laws

fiveprinc

1. **It is simply stated, succinct, and has a clear meaning** - It is imperative that those who enforce and interpret the law, and those who are subject to the law, are able to understand both the letter and the intent of the law.
2. **It is completely successful in achieving its objective** - Every law in a democracy has a problem-solving purpose, or objective, that serves the best interests of the people and reflects their highest aspirations. The ideal law is completely successful in attaining its objective.
3. **It interacts synergistically with other laws** - Laws often have an effect upon, and are affected by, other laws. The ideal law is designed so that its interaction with other laws is synergistic in the attainment of its problem-solving objective.
4. **It produces no harmful side effects** - All human-made products, including laws, have unintended side effects that may be beneficial, neutral, or detrimental. A law that accomplishes its problem-solving goal is not acceptable if its unintended side effects degrade the established living standards or quality of life of the people, or infringe upon human rights. Therefore, the ideal law produces no detrimental side effects upon the human rights, living standards, or quality of life of the people.
5. **It imposes the least possible burdens on the people** - The ideal law imposes the least possible costs and other burdens upon the people so that the maximum positive net benefit of its enforcement is attained. It is cost-efficient, safe, non-intrusive, and user friendly.

ra

First version

Article 1 – This law repeals (cancels) all previuos laws on tax fraud.

Second version

Article 1 – This law repeals Law 320/2005 in the aspects related to tax fraud.

Third version

Article 1 - This law repeals Law 320/2005, whole Art. 1 and Art 7, commas 1 and 3.

dimoflawsvslawfrcons

Context → Quality dimension	Single law	Country legal framework	Federation of Countries Legal framework
(Referential) accuracy		x	x
Clarity	x	x	x
Simplicity	x	x	x
Coherence		x	x
Accessibility	x	x	x
Unambiguity	x	x	x
Conciseness	x	x	x
Global quality index	x		
Level of integration		x	x

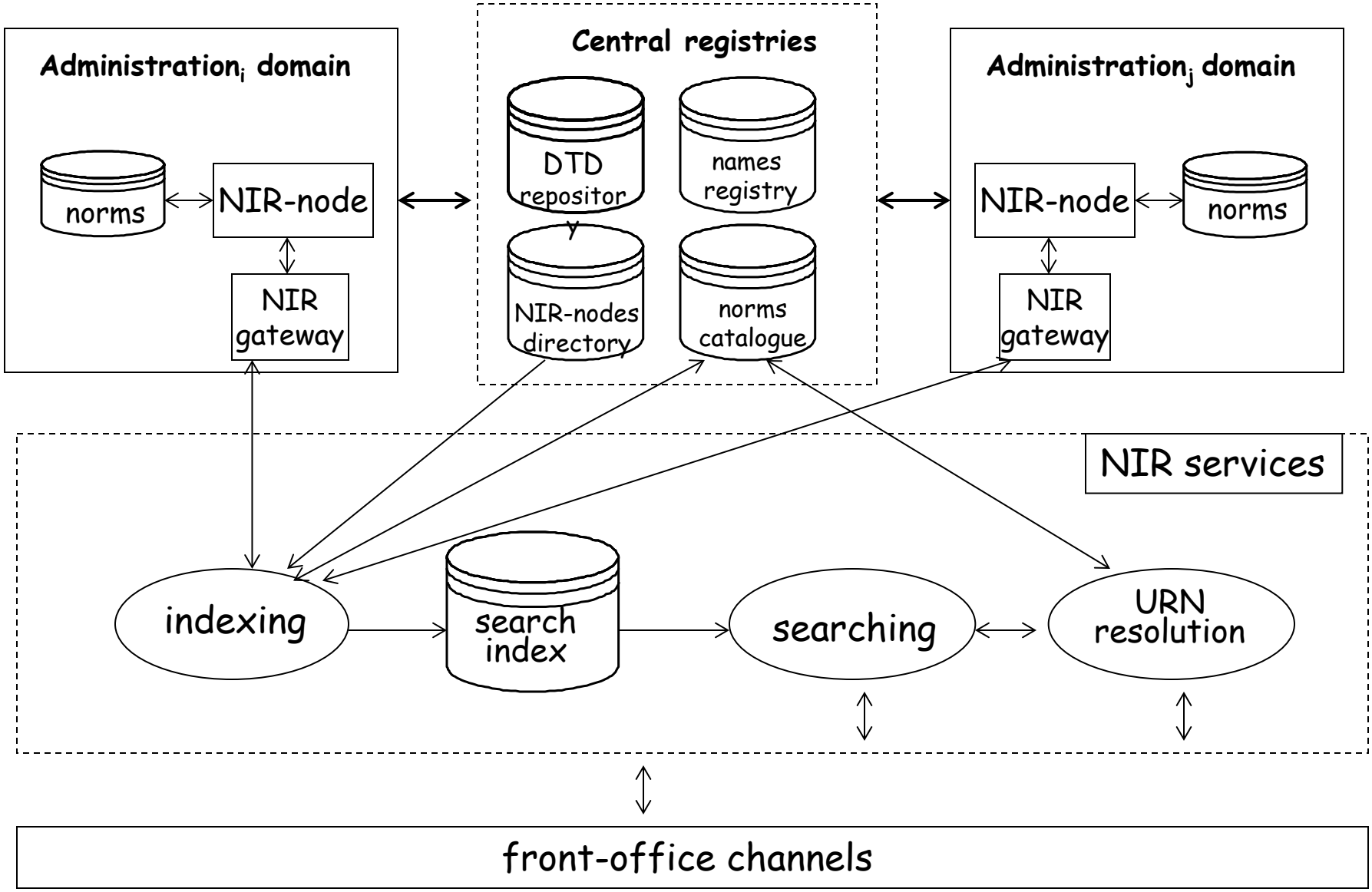
ldvsc

Cluster	Quality dimension	Single law	Legal framework of a single country	Legal framework of a set of federated countries
Accuracy	Referential accuracy Unambiguity	Not relevant x	x x	x x
Redundancy	Conciseness	x	x	x
Readability	Clarity Simplicity	x x	x x	x x
Accessibility	Cultural accessibility	x	x	x
Consistency	Consistency Coherence	Weakly relevant Weakly relevant	x	x
Global quality index		x	Not defined	Not defined

cccmodel

	Correspondence	Consistency	Correctness
A. Text type	a. Appropriateness	2. Purity of genre	3. Application of genre rules
B. Content	4. Sufficient information	5. Agreement between facts	6. Correctness of facts
C. Structure	7. Sufficient coherence	8. Consistent structure	9. Correct linking words
D. Wording	10. Appropriate wording	11. Unity of style	12. Correct syntax and choice of words
E. Presentation	13. Appropriate lay-out	14. Layout adapted to text	15 Correct spelling and punctuation

nirarch


Administration$_i$ domain
norms
NIR-node
NIR gateway
Central registries
DTD repository
names registry
NIR-nodes directory
norms catalogue
Administration$_j$ domain
NIR-node
norms
NIR gateway
NIR services
indexing
search index
searching
URN resolution
front-office channels

Quality dimension	Example of rule
(reference) Accuracy	First version: Article 1 – This law repeals all previous laws on tax fraud Second version: Article 1 - This law repeals Law 122/2005 in the aspects related to tax fraud Third version: Article 1 - This law repeals Law 122/2005, whole Art. 1 and Art. 7, paragraphs 1 and 3
Unambiguity	- Do not use "and/or". Use “or” to mean any one or more. - Use “the” if the reference is unambiguous. Otherwise, use “this”, “that”, “these” or “those”.
Conciseness	- Omit needless language. If a word has the same meaning as a phrase, use the word. - Use the shortest sentence that conveys the intended meaning - Administrative bodies should not use the phrase “in substantially the following form” or “substantially as follows”, since the meaning of “substantially” is ambiguous.
Clarity	- (→) from pertinence) A statement of purpose or occasional example may, however, be helpful to users, including courts interpreting the act. - A suggested order of arrangement of a bill is: a. Short title. b. Preamble; findings; purpose. c. Definitions. d. Scope, exceptions, and exclusions, if any. e. Creation of an agency or office. f. Administration and procedural provisions. g. Substance (state positive requirements in order of time, importance, or other logical sequence). h. Prohibitions and penalties. i. Repeals. j. Saving and transitional provisions to existing relationships, if any. k. Effective dates.
Simplicity	- Select short, familiar words and phrases that best express the intended meaning according to common and approved usage. The language should be dignified, not pompous. Examples: Use “after”, instead of "subsequent to"; use "before" instead of "prior to".
Cohesion & Coherence	- Do not use both a word and its synonym. - Be consistent in the use of language throughout the bill. Do not use the same word or phrase to convey different meanings. Do not use different language to convey the same meaning

Factor ID	Quality factor (scope of the rule) [CRT 2003, OLI 2007]
R1	Use of Abbreviations (as few abbreviations as possible, capita letters, extended expression in glossary)
R2	Number formatting (latin figures, except for measures and percentages)
R3	Date formatting (dd month yyyy hour)
R4	Partition referencing (compliancy to the section hierarchy)
R5	Conventions on units of measurements (written in full, ISO standard)
R6	Act referencing (descending order, full absolute references)
R7	Reference and quotation drafting
R8	Excpetions to rule quotation
R9	Act partitioning
R10	Act partitioning: Sectioning
R11	Act partitioning: Numbering (Latin figures, i.e., letters)
R12	Item drafting
R13	Item numbering
R14	Paragraph drafting and numbering (Arabic figures, no line returns)
R15	Paragraph partitioning
R16	Reference to items or paragraphs
R17	Drafting of modification rules
R18	Numbering of additional items
R19	Numbering of additional paragraphs
R20	Additional letters and numbers (paragraph partitions)
R21	Expression of the definitive rule proper

qf

Qfvsqd - Classification of quality factors wrt quality dimension clusters

Quality factor Qdim Cluster	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	R21
1. Correctness, Accuracy, Precision		X	X		X	X	X	X			X		X	X	X	X		X	X	X	X
2: Completeness, Pertinence			X	X		X	X	X									X				
3: Minimality, Redundancy, Compactness				X			X	X			X		X	X	X	X		X	X	X	
4: Consistency, Coherence, Compliance	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
5: Readability, Comprehensibility, Usability	X			X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
6: Accessibility			X	X		X			X	X	X	X	X	X	X	X		X	X	X	X

1
2: Completeness, Pertinence
3: Minimality, Redundancy, Compactness
4: Consistency, Coherence, Compliance
5: Readability, Comprehensibility, Usability
6: Accessibility

Figure dimensionsvstypesofinfo

Quality Dimension Cluster	Structured data	Geographic Maps	Images	Unstructured Texts	Laws and legal frameworks
Correctness/ Accuracy/ Precision	Schema w.r.t requirements w.r.t. the model Instance Syntactic Semantic Domain dependendent (ex. Last Names, etc.)	Instance Spatial accuracy - Relative/Absolute - Relative Inter layer - Locally increased r.a. - External/Internal - Neighbourhood a. - Vertical/Horizontal/Height Attribute accuracy. Domain dependent accuracy (ex. Traffiic at critical inters, Urban vs rural areas, etc.) Acc. of raster represntation	Accuracy Syntactic Semantic "Reduced" semanic Genuineness Fidelity Naturalness	Accuracy Syntactic Semantic Structural similarity	Accuracy Precision Objectivity Integrity Correctness Reference accuracy
Completeness/ Pertinence	Schema Completness Pertinence Instance Value C., Tuple C., Column C., Relation C., Database C.	Completeness (btw different datasets) Pertinence	Completeness	Completeness	Objectivity Completeness
Temporal	Currency – Timeliness - Volatility	Recency/ Temporal accuracy/ Temporal resolution			
Minimality/ Redundancy/ Compactness/ Cost	Schema Minimality Redundancy	Redundancy	Minimality		For a law: Conciseness For a legal framework: Minimality, Redundancy
Consistency/ Coherence/ Interoperability	Instance Intrarelational Consistency Interrelational Consistency Interoperability	Consistency Object consistency Geometric consist. Topological consist. Interoperability		Cohesion Referential, Temporal, Locational, Causal, Structural Coherence Lexical Nonlexical	Coherence Consistency among laws Consistency among legal frameworks
Readability/ Comprehensibility/Usability/ Usefullness Intpretability	Schema Diagrammatic Readability Compactness Normalization	Instance Readability/Legibility Clarity Aesthetics	Readability, Usefullness	Readability Comprehensibility	Clarity Simplicity
Accessibility…	Instance Technological Channel Physical (W3C)	Instance Privacy	Physical Accessibility (W3C)	Cultural Accessibility	Accessibility of the consolidated Act on a given domain
Others	Lineage	Effectiveness Lineage Adaptation			Effectiveness, Transaprency Usefullness Applicability, Accountability

Cartographic Generalization, lies and truth
tradeoff between Accuracy and Readability

Different combinations, amounts of application, and different orderings of these techniques can produce different yet aesthetically acceptable solutions. The focus is not on making changes to information contained in the database, but to solely focus upon avoiding ambiguity in the interpretation of the image.

The process is one of compromise reflecting the long held view among cartographers that making maps involves telling small lies in order to tell the truth!

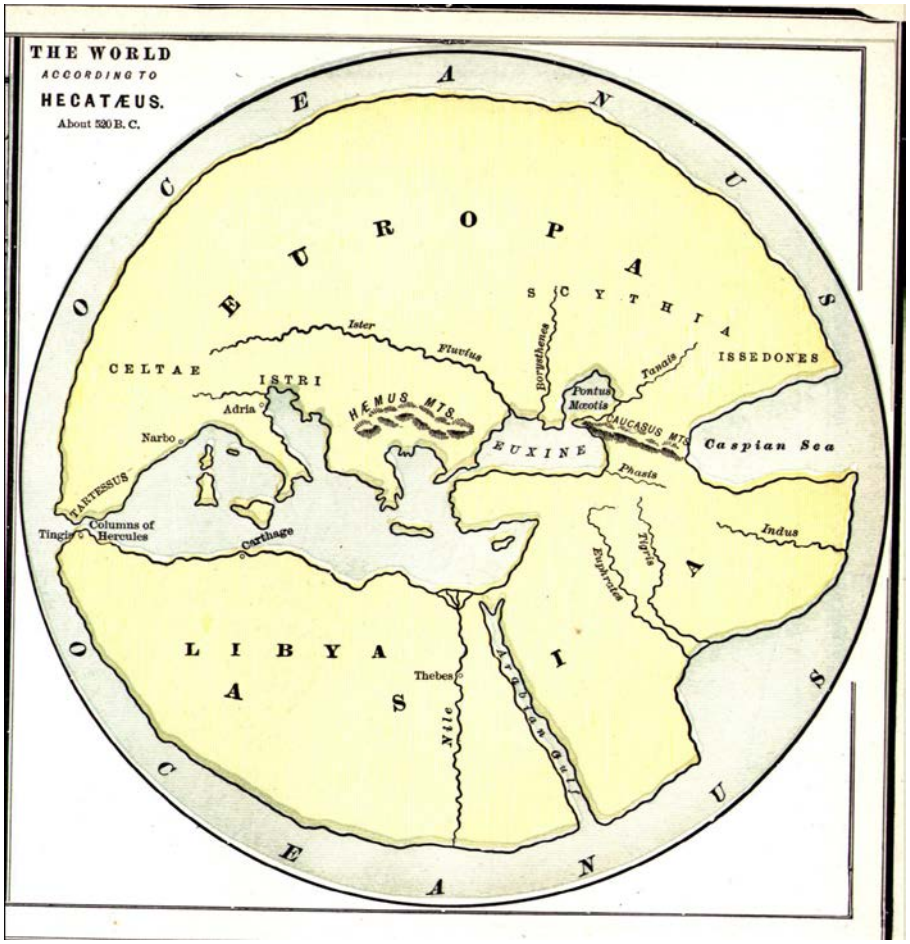
Information quality dimensions

a. From data quality to information quality
   a. The two coordinates of the problem: type of information vs domain of information
   b. Chapter organization and contents
b. Quality of maps
   a. Definition and conceptual structure of maps
   b. Map quality dimensions vs map conceptual structure
   c. Accuracy
   d. Up-to-dateness
   e. Correctness
   f. Completeness
   g. Consistency
   h. Quality of abstractions and quality of maps
c. Quality of loosely structured texts
   a. Readability
   b. Comprehensibility
   c. Coherence
   d. Cohesion
d. Quality of laws
   a. Quality dimensions in laws, legal frameworks, federations of legal frameworks
   b. (Referential) accuracy
   c. Clarity
   d. Simplicity
   e. Coherence
   f. Accessibility
   g. Unambiguity
   h. Conciseness
   i. Global quality index

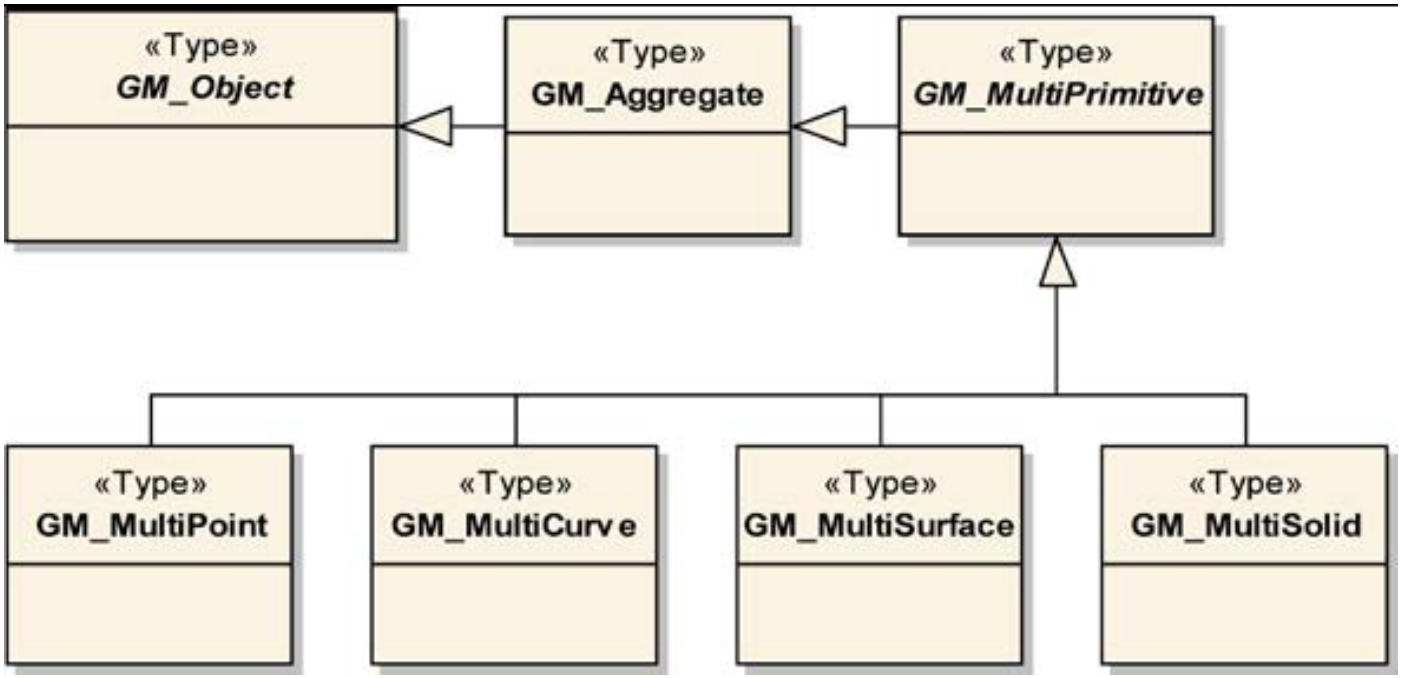
Hecateus Map (520 B.C.)

Hecateus Map (520 B.C.)

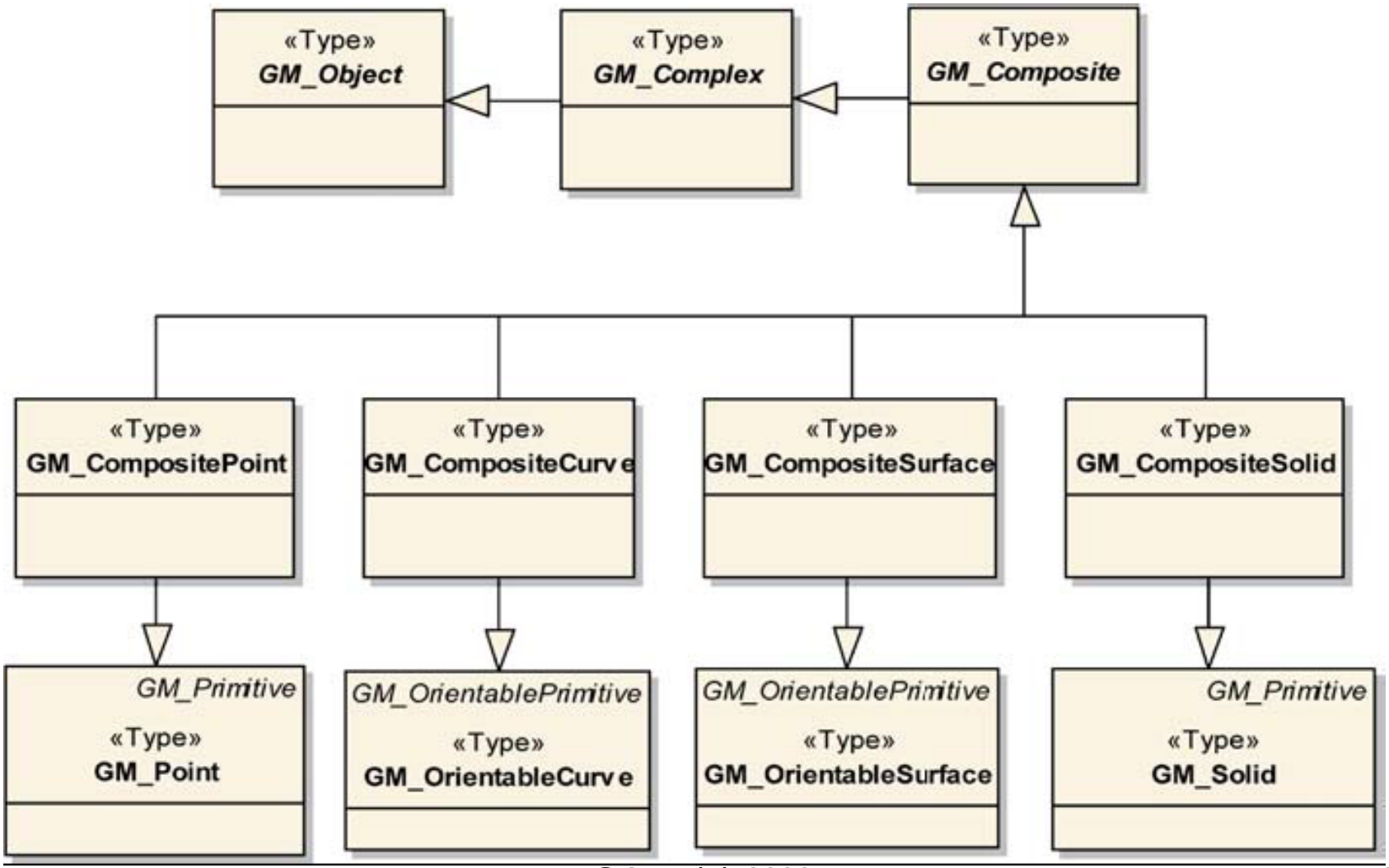
A representation, usually on a flat surface, as of the features of an area of the earth or a portion of the heavens, showing them in their respective forms, sizes, and relationships according to some convention of representation



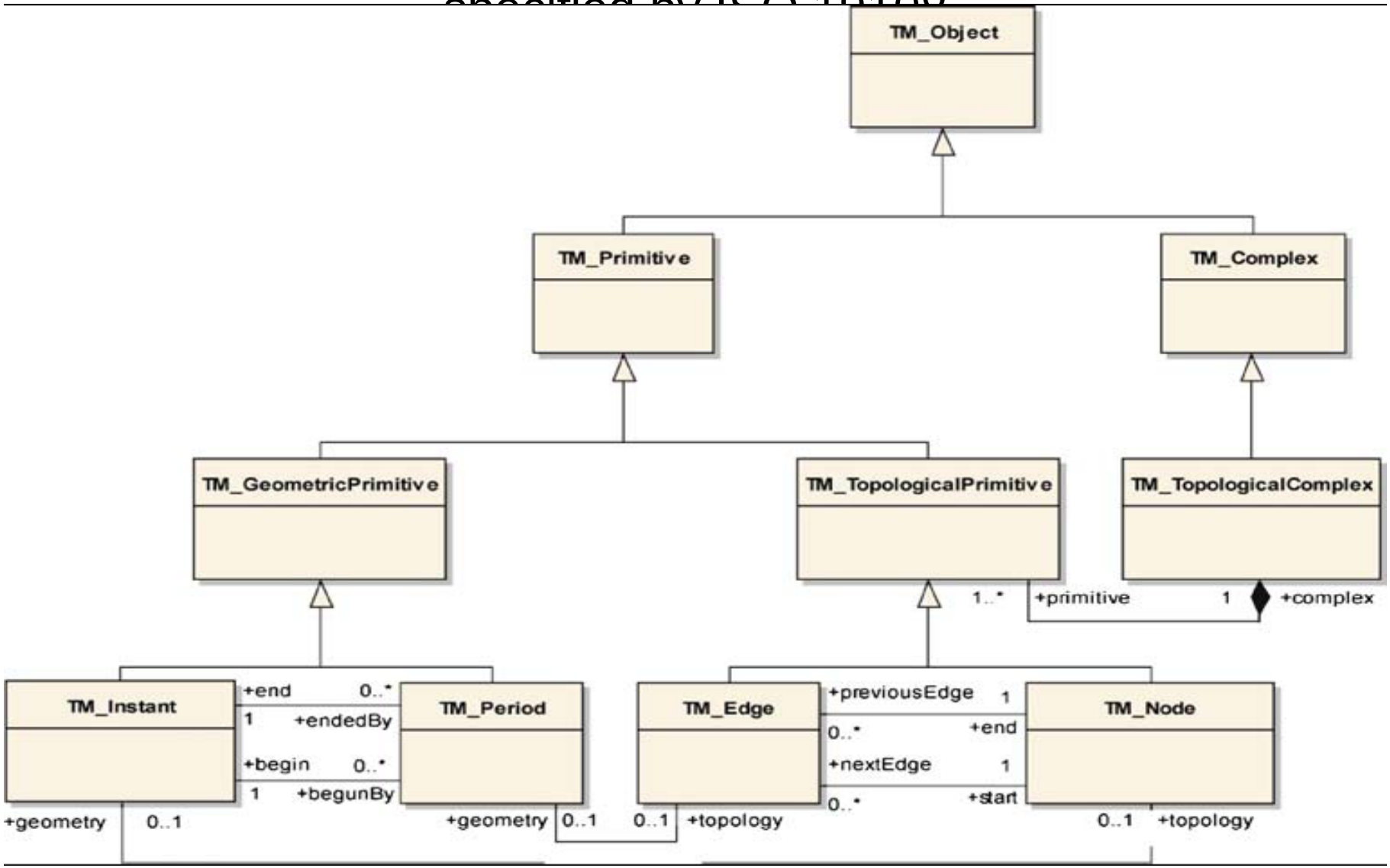
Modeling with ISO 191xx Standards - 2
Aggregate geometries specified by ISO 19107



Modeling with ISO 191xx Standards - 3
Aggregate geometries specified by ISO 19107



Modeling with ISO 191xx Standards **- 4**
Temporal primitives, topological primitives and complex specified by ISO 19108



Qdim of unstructured text

- Readability
- Understandability
- Cohesion
- Coherence

Levels of discourse in \cite{graesser2011computational}

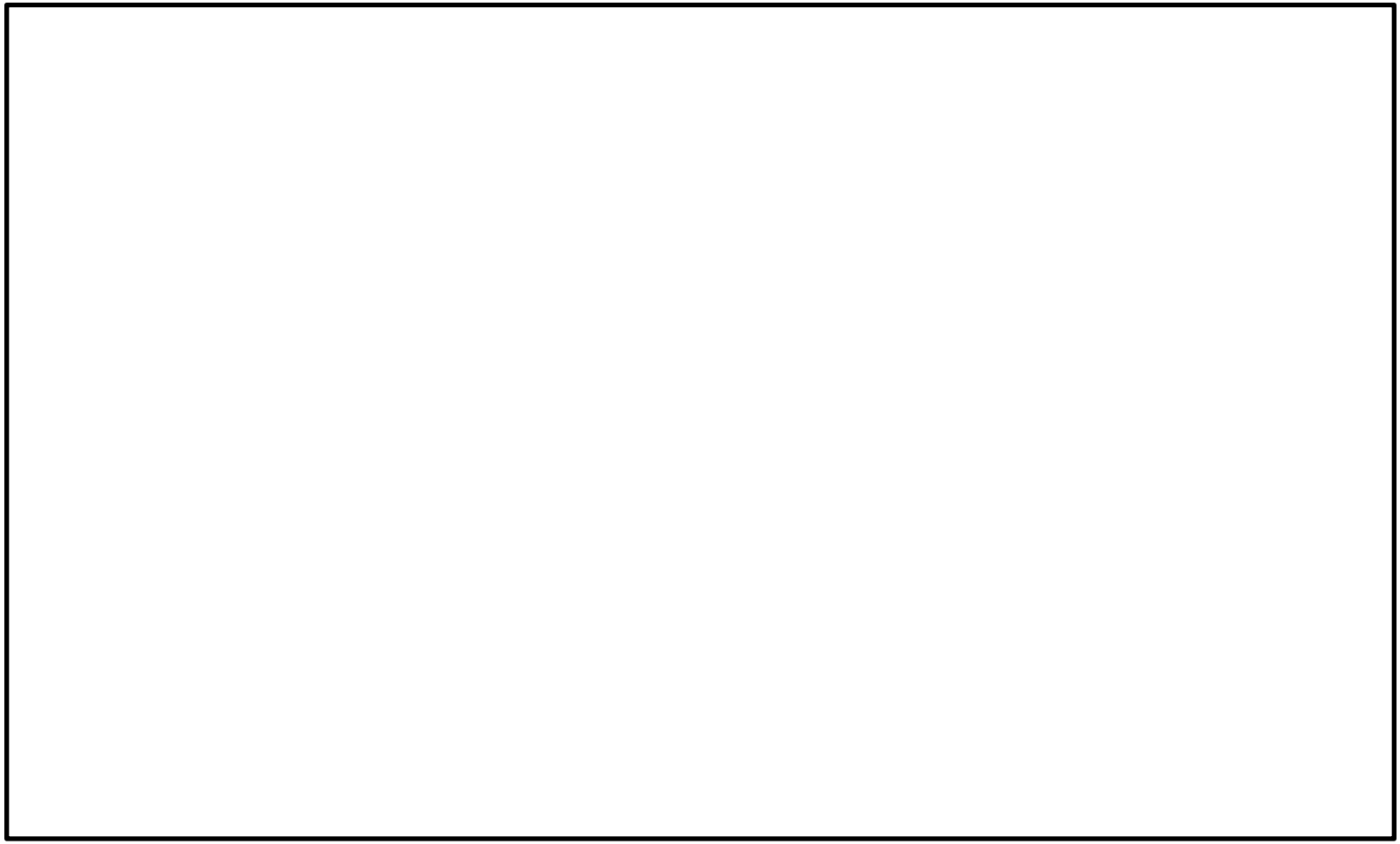
(1) Surface code
- Word composition (graphemes, phonemes, syllables, morphemes, lemmas, tense, aspect)
- Words (lexical items)
- Part of speech categories (noun, verb, adjective, adverb, determiner, connective)
- Syntactic composition (noun-phrase, verb-phrase, prepositional phrases, clause)
- Linguistic style and dialect

(2) Textbase
- Explicit propositions
- Referents linked to referring expressions
- Connectives that explicitly link clauses
- Constituents in the discourse focus versus linguistic presuppositions

(3) Situation model
- Agents, objects, and abstract entities
- Dimensions of temporality, spatiality, causality, intentionality
- Inferences that bridge and elaborate ideas
- Given versus new information
- Images and mental simulations of events
- Mental models of the situation

(4) Genre and rhetorical structure
- Discourse category (narrative, persuasive, expository, descriptive)
- Rhetorical composition (plot structure, claim + evidence, problem + solution, etc.)
- Epistemological status of propositions and clauses (claim, evidence, warrant, hypothesis)
- Speech act categories (assertion, question, command, promise, indirect request, greeting, expressive evaluation)
- Theme, moral, or point of discourse

(5) Pragmatic communication
- Goals of speaker / writer and listener / reader
- Attitudes (humor, sarcasm, eulogy, deprecation)
- Requests for clarification and backchannel feedback (spoken only)



Feature set in \cite{aluisio2010readability}

Cognitively motivated features - Basic
Number of words
Number of sentences
Number of paragraphs
Average number of words per sentence
Average number of sentences per paragraph
Average number od syllabes per word
Cognitively motivated features - complex
Incidence of functional words
Average number of verb hyperonyms
Number of persoan pronouns
Number of negations
Number of connectives
Verb ambiguity ratio
Boun ambiguity ratio
Syntactic constructions considered in the text simplification system
Incidence of clouses
Incidence of relative clauses
Incidence of subordination
Features derived from n-gram language models (LM) plus out-of-vocabulary rate scores
LM probability of unirams
LM probability of bigrams
LM probability of trigrams
Out-of-vocabulary words

Cohesion and Coherence

- Although related to the problem of measuring the difficulty of a text, readability formulas *cannot* capture text cohesion or coherence.

- Both cohesion and coherence represent how words and concepts conveyed in a text are *connected* on particular levels of language, discourse and world knowledge.

- Cohesion is a characteristic of the text → measurable, objective
- Coherence is a characteristic of the reader's mental representation → subjective

Cohesion and Coherence

- Cohesion is an *objective property* of the explicit language and text. There are explicit linguistic devices that allow to express connections (relations) between words, sentences etc. These cohesive devices cue the reader on how to form a coherent representation.
- Coherence results from an *interaction between text cohesion and the reader*. The coherence relations are constructed in the mind of the reader and depend on the skills and knowledge that the reader brings to the situation.

  A particular level of cohesion may lead to a coherent mental representation from one reader but an incoherent representation for another.

  [Graesser, McNamara, Louwerse and Cai, 2004].

Cohesion and Coherence

- The literature distinguishes various kinds of cohesion and coherence. One common distinction is between *local* and *global* levels. Both cohesion and coherence are locally and globally structured. The reader finds local cohesion relations between adjacent clauses in the text and global cohesion links between group of clauses and groups of paragraphs.

- Moreover, the following conceptual categories of cohesion and coherence can be distinguished: *referential*, *temporal*, *locational*, *causal* and *structural*.

Cohesion

- Halliday and Hasan (1976): “The use of certain linguistic devices to link or tie together textual units” (the phenomenon where the interpretation of some element of discourse depends on the interpretation of another element and the presupposing element cannot be effectively decoded without recourse to the presupposed element).
- Lexical cohesion (lexical relations):
  - Use of lexical relations between words in the two units, such as **identical word (reiteration)**, **synonymy**, **hyperonimy, conjunction:**
    - Before winter **I** built a chimney, and **shingled** the sides of my **house.**
    - **I** thus have a tight **shingled** and plastered **house**.
    - Peel, core and slice **the pears and the apples**. Add **the fruit** to the skillet.
- Nonlexical cohesion
  - Use of non lexical relations such as anaphora
    - The **Woodhouses** were first in consequence there. All looked up to **them**.
- Cohesion chain:
  - Peel, core and slice **the pears and the apples**. Add **the fruit** to the skillet. When **they** are soft…

Coherence

- Coherence is a key concept of text linguistics.

•Coherence is especially relevant to the research on text comprehension I6 APPL: authors should design a text in such a way that the addressee may detect the relationships linking individual text constituents and thus may build a coherent mental model of the text's content.

- Coherence has a *subjective metric.*

Coherence: example

- Text coherence can be reconstructed using a set of *coherence relations,* which relate the semantic constituents of a text to one another.
- The type of the relation is either made explicit, by means of *connectives* (example 1 below).
- Or the relation remains implicit and thus has to be inferred, via context clues and background knowledge, by the reader (example 2).
- To reconstruct the relation between individual constituents, such inference must often be based on quite complex *frame and script knowledge* – in example (3) knowledge of birthday parties and piggybanks.

(1) As she is sick, Jennifer stays home.

(2) Jennifer is sick. She stays home.

(3) Jane was invited to Jack's birthday party. She wondered if he would like a kite. She went to her room and shook her piggy bank. It made no sound.

Coherence

Since many coherence relations remain implicit and must be interpreted by the reader, one cannot determine the coherence structure of a text in a straightforward fashion. There are two main perspectives:

- From the perspective of *discourse production*, coherence is a property of the mental representation of the content that the text composition is to convey. This property may be reconstructed as the *author`s coherence structure*. This structure determines the author's strategies for composing the text and is reflected in the surface text by means of *coherence cues*.

- These coherence cues, in turn, support the text recipient in building a coherent, mental model of the text content. Thus, from the perspective of *discourse comprehension*, coherence is a property of the mental representation that is built while reading the text. This property may be reconstructed as the *reader`s coherence structure.*
→ I6 APPL

ra

First version

Article 1 – This law repeals (cancels) all previuos laws on tax fraud.

Second version

Article 1 – This law repeals Law 320/2005 in the aspects related to tax fraud.

Third version

Article 1 - This law repeals Law 320/2005, whole Art. 1 and Art 7, paragraphs 1 and 3.

Global quality index

The Working Group of the Regional Council of Tuscany (flanked by the Italian Interregional Law Observatory) has proposed an

- Index of Legislative Quality [OLI2007, CRT2002]
  - 100% highest quality
  - 0% lowest quality

- Quality is intended as a measure of “how well the legislative text complies to the legislative drafting rules”.
  - which rules?

- Rules that, for their technical nature, can be applied directly by the regional legislative offices with no need for further interpretation.

- These rules regard the syntactic and structure-related level of the text
  - e.g., formatting constraints, naming and referencing conventions, domain-specific expressions and terms.

- The Working Group identified 21 rules
  - extracted from the 93 items contained in the “Rules and Suggestions for Legislative Drafting” (1991, 2002, 2007)
  - denoted as “quality factors”.

- Assumptions on the quality factors:
  - Inter-independency.
  - Equal intrinsic relevance.
  - The more frequently the rule is applied properly (within the whole reference body of laws), the higher is the factor's contribution to the overall quality of the single law.
    - Factors can be weighted according to the frequency of proper compliance of the corresponding rule.
    - Weights (w_i) from 5 (very relevant) to 1 (not relevant).
      - Therefore, these weights are relative to a specific time period, Region and legislative scope.

- Methodology of Assessement
  - For each law it is calculated:
    - a Qualitative Standard (QS)

$$QS = \sum_i W_i$$ with i: R_i rule is properly applied within the law

    - a Qualitative Profile (QP)

$$QP = QS - \sum_i W_i$$ with i: R_i rule is not (properly) applied within the law

    - a Quality Index (QI)

$$QI = QP / QS$$

    - an Improvement Index (II)

$$II = 1 - QI$$

- Legislative Quality Assessment
  - The Quality Index can be used to identify laws and sections of laws to be syntactically amended/replaced.
  - Other indicators (e.g., Flesh Index, Gulpease Index) can be used to identify laws and sections to make more readible.

- Legislative Quality Improvement
  - Quality-oriented policies from above.
  - Comprehensive and unique set of drafting rules.
  - Training programmes of stakeholders involved.
  - Drafting Editors
    - with macros, automatic checkers and DTD validators.
  - Legislative DBMS and Law URIs (cf. NIR project)
    - for automatic and sound cross-referencing.

Data quality in maps

Given the significant number of sources for the road map data and the heterogeneity across the data, it became necessary to define data quality in the context of digital road maps. Data quality refers to the relative accuracy and precision of a particular road map database.

The purpose of the data quality report is to provide adequate information to the users to evaluate the fitness of the data for a specific use. There are several map accuracy standards, including the well-known National Map

Accuracy Standard (NMAS) and the American Society for Photogrammetry and Remote Sensing (ASPRS) standard [1]. The standards consist of four components namely:

1. Lineage: This component deals with the narrative of the source materials used and procedures adopted to build the product.
2. Positional Accuracy: This defines the error in position of features. In digital road maps, this component is the most critical.
3. Attribute Accuracy: This represents the expected error in attributes such as road names.
4. Completeness: This defines the fraction of the realworld features represented on a map.

In addition, topological consistency is of concern for digital road maps in the context of navigation systems to facilitate graph computations such as shortest path algorithms.

Quality dimensions of maps

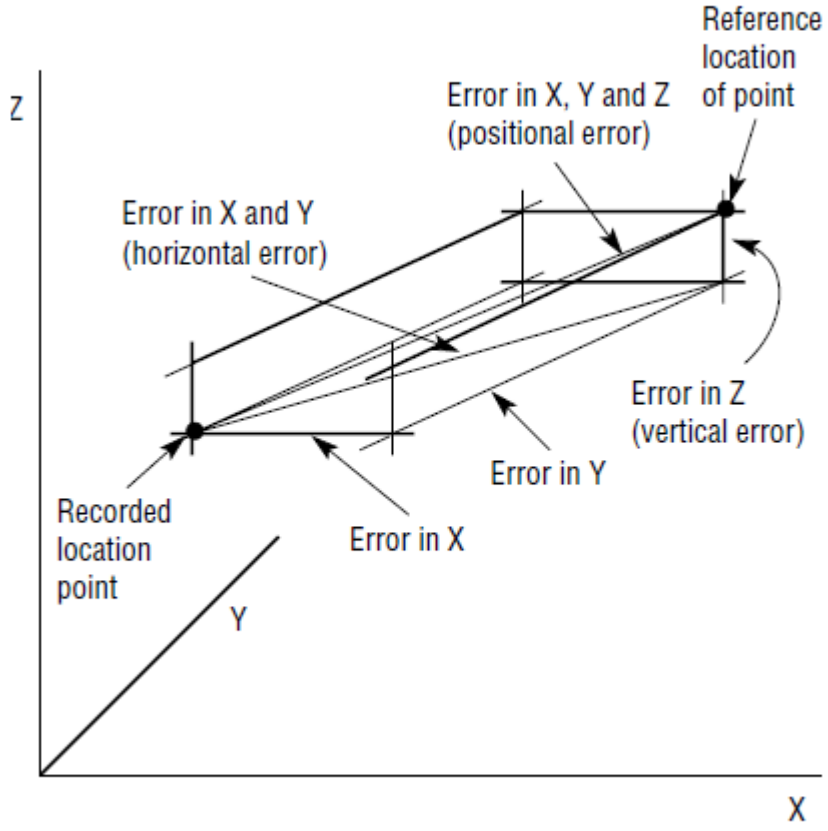


Fig 2. Measuring components of spatial error.

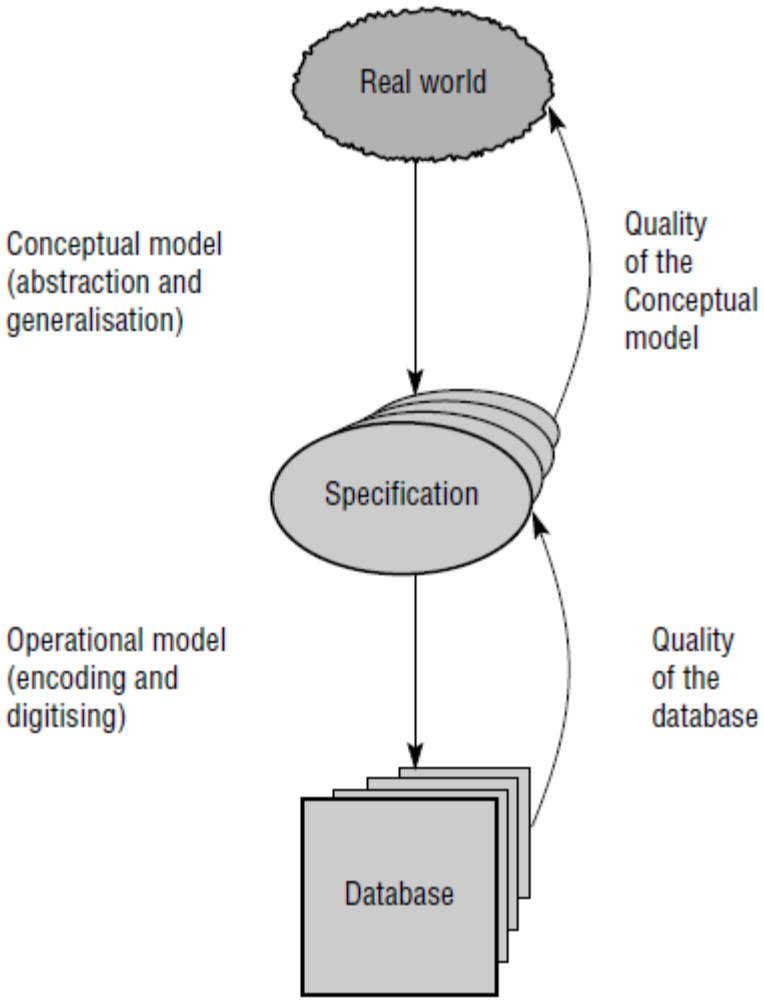


Fig 1. The mediating role of the database specification in assessing data quality.

Table 1 Data quality components in SDTS.

Component	*Description*
Lineage	Refers to source materials, methods of derivation and transformations applied to a database. ● Includes temporal information (date that the information refers to on the ground). ● Intended to be precise enough to identify the sources of individual objects (i.e. if a database was derived from different source, lineage information is to be assigned as an additional attribute of objects or as a spatial overlay).
Positional accuracy	Refers to the accuracy of the spatial component. ● Subdivided into horizontal and vertical accuracy elements. ● Assessment methods are based on comparison to source, comparison to a standard of higher accuracy, deductive estimates or internal evidence. ● Variations in accuracy can be reported as quality overlays or additional attributes.
Attribute accuracy	Refers to the accuracy of the thematic component. ● Specific tests vary as a function of measurement scale. ● Assessment methods are based on deductive estimates, sampling or map overlay.
Logical consistency	Refers to the fidelity of the relationships encoded in the database. ● Includes tests of valid values for attributes, and identification of topological inconsistencies based on graphical or specific topological tests.
Completeness	Refers to the relationship between database objects and the abstract universe of all such objects. ● Includes selection criteria, definitions and other mapping rules used to create the database.

FAO Interdisciplinary Database: Spatial Standards and Norms

FAO q d classification

- Accuracy
  - Spatial coordinate
    - Horzontal
    - Vertical
  - Topology
    - Relative
    - Absolute
- Completeness
- Correctness
- timeliness
- Integrity (or internal consistency)
- Graphic quality

Roads

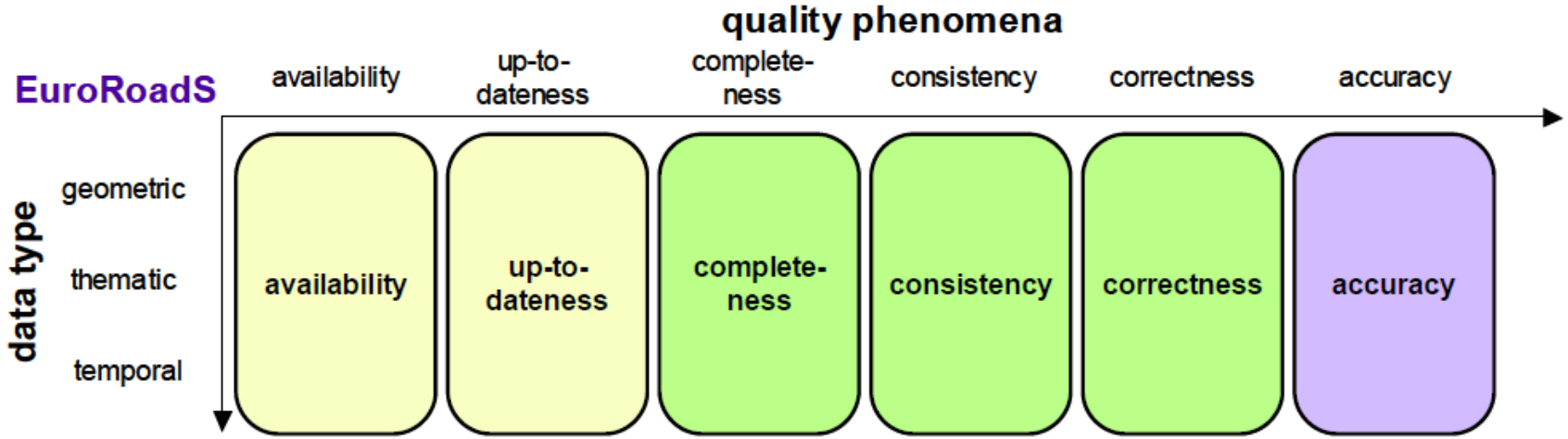


Figure 3: Quality characteristics in the EuroRoadS quality model

- accuracy
  - Geo
    - absolute position accuracy
    - quantitative attribute accuracy
  - Them
  - Temp
- dependability
  - Availability
    - Failure rate
  - Up-to-dateness
    - Last update
    - Rate of change
    - Temporal lapse
- integrity
  - correctness
    - Geometric
    - Topological
    - thematic
  - completeness,
  - consistency,
    - thematic consistency
    - topological consistency
    - geometric consistency
- Correctness
    - Topological

Table 2: Definitions of quality characteristics and quality parameters for the EuroRoadS quality model

groups of quality characteristics	quality characteristics	definition	possible quality parameter
dependability	availability	degree to which geographic data are available at a certain place and at a defined time	failure rate
			user-defined
	up-to-dateness	degree of adherence of geographic data to the time changing universe of discourse	last update
			rate of change
			temporal lapse
			user-defined
integrity	completeness	degree of adherence of the entirety of geographic data (features, their attributes, and relationships) to the entirety of the universe of discourse	omission (*)
			commission (*)
			user-defined
	correctness	degree of adherence of existence of geographic data (feature(s), attributes, functions, relationships) to corresponding elements of the universe of discourse, up-to-dateness being presumed	geometric correctness
			topological correctness
			thematic correctness
			user-defined
	consistency	degree of adherence of geographic data (data structure, their features, attributes, and relationships) to the models and schemas (conceptual model, conceptual schema, application schema, and data model)	geometric consistency
			topological consistency (*)
			thematic consistency
			user-defined
accuracy	accuracy	degree of adherence of geographic data to the most plausible resp. true value.	absolute position accuracy (*)
			relative position accuracy (*)
			quantitative attribute accuracy (*)
			temporal accuracy of time measurement (*)
			user-defined

Road network quality

- accuracy
  - absolute position accuracy
  - quantitative attribute accuracy
- correctness
  - topological
- dependability
  - availability
  - Up-to-dateness
- integrity –
  - completeness,
  - consistency,
    - thematic consistency
    - topological consistency
    - geometric consistency
- Correctness
    - Topological

Veregin geospatial quality dim.

- Accuracy
  - Temporal
  - Thematic
- Precision (or resolution)
  - Spatial resolution
  - Temporal
  - Thematic
- Consistency
  - Logical
  - topological
- Completeness

	Space	Time	Theme
Accuracy			
Precision			
Consistency			
Completeness			

According to some critics, technologies such as GIS have led to the ascendance of a new geospatial science focused on the goal of producing ultimately truthful and objective representations of reality. This goal is seen as a byproduct of the new technological means with its appeals to neo-positivism, reductionism, instrumentalist thinking, and naïve empiricism in which ‘reality’ is uncontested and objectively measurable (e.g. Harley 1991;Wood 1992). According to this view, producers of geospatial databases make no allowance for the possibility that these databases embed specific social and institutional values. As such, GIS promulgates the myth of an objective science which always produces the best delineations of reality (Harley 1989).

While there is some foundation to this critique, it would be unfair to suggest that producers of geospatial data are unaware of the limitations of these data. Like their manually-produced map counterparts, geospatial data are not intended to be miniature replicas of ‘reality’. Rather they emphasise some aspects of the environment and suppress others in an effort to convey a particular message (Martin, Chapter 6; Raper, Chapter 5).What is contained in a database is a function not only of the nature of the external environment but also the values of the society and institution within which the database was constructed (Turnbull 1989). Values are embedded at the modelling stage, where they impact on database content, and at the representation stage where they affect database form.

Values are not always embedded deliberately. Broad social values are often taken for granted and may not be consciously recognised. Hence databases often unintentionally reflect and legitimate the social order (Harley 1989). Broad social values form the backdrop for more specific values that reflect institutional characteristics. Perhaps the most significant of these is institutional mandate, which defines institutional mission for data collection and dissemination. For specific databases, mandate is formalised as a set of design guidelines that outline the rules for data collection, encoding, and representation.

Unlike broad social values, values deriving from institutional mandate can be articulated, documented, and communicated to database consumers through the medium of metadata. This communication process is important since it affect the consumer's understanding of the limitations of database and facilitates its appropriate use. Especially useful in this context is the concept of the 'specification' describing the intended contents of the database. The specification is the reference standard against which the database is compared in order to assess completeness and other data quality components. The specification concept explicitly recognises that each database has a particular set of objectives and that embedded in these objectives is the formal expression of the values associated with institutional factors.

What are the implications for the debate over values? First, geospatial databases are not intended to be accurate mirrors of reality. Rather, they are designed to conform to a database specification which could just as easily be a description of perceived reality. Second, geospatial data producers are generally aware of the significance of values. The database specification is in fact a formal statement of he values that are embedded in a given database. Third, values can be communicated to database consumers who can then use this information to assess the appropriateness of the database for a particular task. Knowledgeable map users have of course always been aware of data limitations.

These are important conclusions since the alternatives are not particularly attractive. For example, some critics have claimed that given the dependence on social values it is not possible to distinguish between competing representations of the same geographical space. Thus it has been argued that the distinction between propaganda and truth is artificial and must be dismantled, as must the arbitrary dualism between art and science (Harley 1989). According to this view, all representations are equally valid since they are all expressions of one's personal values, or the values of one's culture, or the values of one's institution, any one of which has no more claim to legitimacy than any other. This anarchistic epistemology implies that we have no agreed standard of reference and no basis for communicating biases and assumptions. On the other hand, if databases are to be more than just personal artistic diversions and are to convey information rather than simply express the values and viewpoints of their creator, then they must be able to convey their meaning to a broad spectrum of users.

- Data quality criteria define to what standards data must comply in order to be usable within the system. Standards for GIS data will normally depend upon the accuracy required for the datasets. In the GIS environment, accuracy will depend upon the scale at which the data is produced, and at which scale the dataset is meant to be used. Three types of positional accuracy can be distinguished:
- *Relative accuracy*: a measure of the deviation between two objects on a map and is normally described in terms of + or - the number of measurement units the feature is located apart from its correspondent map feature.
- *Absolute accuracy*: evaluates the measure of the maximum deviation between the location of the map feature and its location in the real world.
- *Graphic quality*: refers to the visual cartographic display quality of the data, and pertains to aspects such as the data's legibility on the display, the logical consistency of map graphic representations, and adherence to common graphic standards

- Informational quality: relates to the level of accuracy of both map graphic features and attribute data. There are four basic categories for assessing these qualities:
  - completeness
  - correctness
  - timeliness
  - integrity
- Together, these aspects of informational quality comprise the extent to which the dataset will meet the basic requirements for data conversion acceptance. See more details in Annex E.

Recommended Standards

Horizontal accuracy

Maximum allowable error according to map scale:

Scale 1:1,000,000:	**600m**
Scale 1:5,000,000:	**3,000m**
Scale 1:10,000,000:	**6,000m**
Scale 1:40,000,000:	**24,000m**

Completeness

Allowed percentage of missing features:

Country boundaries: 0%
Thematic layers: 1%

Annex E: Data Quality Standards
Positional Accuracy

- USGS has published the National Map Accuracy Standards (see http://mapping.usgs.gov/standards), which define accuracy for map features at appropriate scales. A summary of the accuracy measures are reported below:
- Decimal-degree (longitude/latitude) coordinates for geographic data should be recorded to a minimum 5 significant digits to the right of the decimal point and stored in double precision attribute or database fields. Any calculations done with location data should be done at double precision with the results rounded or truncated to the appropriate propagated error limits.
- No more than 10 percent of the points tested shall be in error by more than 1/50 inch (~0.51mm), measured at the publication scale. These limits of accuracy shall apply to positions of well-defined points. Well-defined points are those that are easily visible on the ground/map: intersections of roads, railroads or rivers; corners (or center points) of large buildings blocks.
- Vertical accuracy, as applied to contour maps on all publication scales, shall be such that not more than 10 percent of the elevations tested shall be in error by more than one-half the contour interval. In checking elevations taken from the map, the apparent vertical error may be decreased by assuming a horizontal displacement within the permissible horizontal error for a map of that scale.

- ***Completeness*** - Completeness is an assessment of the dataset's existing features against what should currently be located within the dataset. Completeness will also relate to the attribute data to assess whether all of the necessary attributes are accounted for. A typical requirement for the bottom limit of dataset completeness is that not more than 1% of the features and attributes existing in the source data will be missing in the output.
- ***Correctness*** - Correctness relates to knowledge of the information contained in a dataset. A map accurately located feature which is incorrectly coded has a problem with correctness. The bottom limit of correctness for acceptance should be set for individual datasets. It is not acceptable, in fact, to have even a single error in the coding of the country boundary polygons, while a limited number of errors in soil coverages might be tolerated.
- ***Timeliness*** - Timeliness is based upon the “currency” of a dataset, providing information on how up-to-date it is and on its expiry date. Timeliness is not dealt with in this report.

- ***Integrity*** - The integrity of a dataset is a measure of its internal consistency. Database integrity means that a dataset maintains its connectivity and topological consistency, which means, lines properly connected, absence of overshoots or undershoots, polygons are closed and labeled with absence of errors in the labels. In order to maintain database integrity, missing or duplicate records or features should be avoided. The five points listed below are guidelines provided by the UN Cartographic Section for maintaining database integrity with respect to the preservation of topology.
- Arc/Info coverages can be generated, have attributes attached and have topology created
- Features are split at international boundaries
- There are no coincident lines within a single coverage
- There are no intersections within a line coverage
- All polygons have labels apart from the universal polygon