Quality-Aware Query Processing

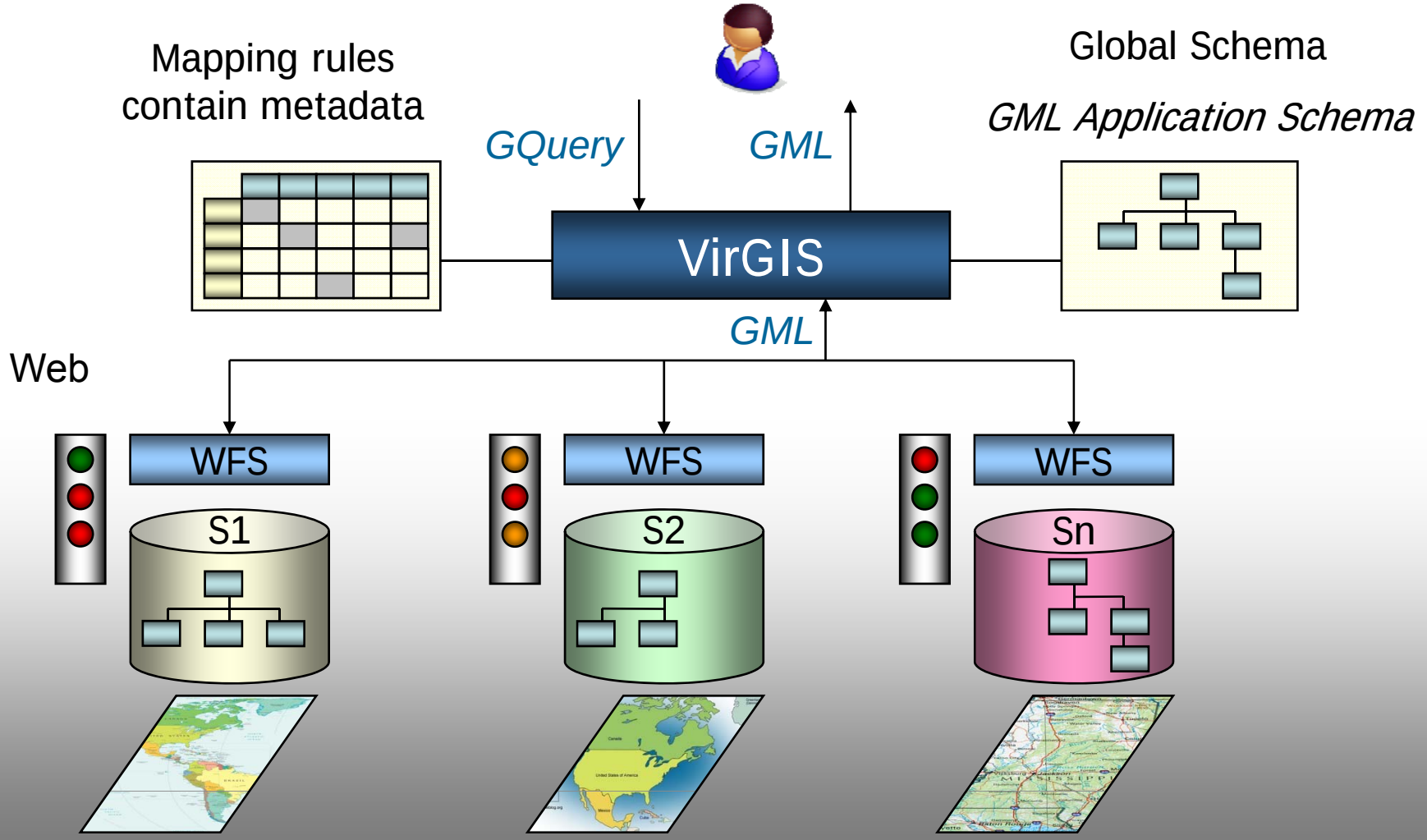
Omar Boucelma

LSIS, Université Paul Cézanne, Aix-Marseille

Credit: M. Essid, Y. Lassoued

Kick-off Meeting STIC-AMSUD
Recife, July 22nd 2008

VirGIS Mediation System



Metadata Model

- Metadata Standard
  - ISO-19115 (specification)
  - ISO-19139 (XML implementation)
- Profile
  - Subset of the discovery metadata specified by the European Committee for Standardization (CEN)

  +
  - Subset of the quality parameters of ISO-19115

Facilitate discovery & define fitness for use

Discovery Metadata

Identification Information		Information to uniquely identify the data
Language	L	Language used within the dataset
Citation		Citation data for the resource
Dataset Title	DT	Name by which the resource is known
Dataset - Publication	DP	Date of publication of the resource
Extent Information		Information about horizontal, vertical and temporal extent
Geographic Extent		Geographic Area of the dataset
Geographic Bounding Box	BB	Geographic position of the dataset
Vertical Extent	VE	Vertical domain of the dataset
Temporal Extent	TE	Time period covered by the content of the dataset
Point of Contact		Identification of and means of communication with, person(s) and organization(s) associated with the resource
Organization Name	*[R]*-ON	Name of responsible organization
Role	[R]	Function performed by the responsible party (Res. Provider [Pvd],
Reference system Information		Information about the reference system
Reference system Identifier		Identifier used for reference systems
Code	RSC	Alphanumeric value identifying an instance in the namespace
Code Space	RSCS	Name or identifier of the person or organization responsible for namespace

Data Quality Metadata

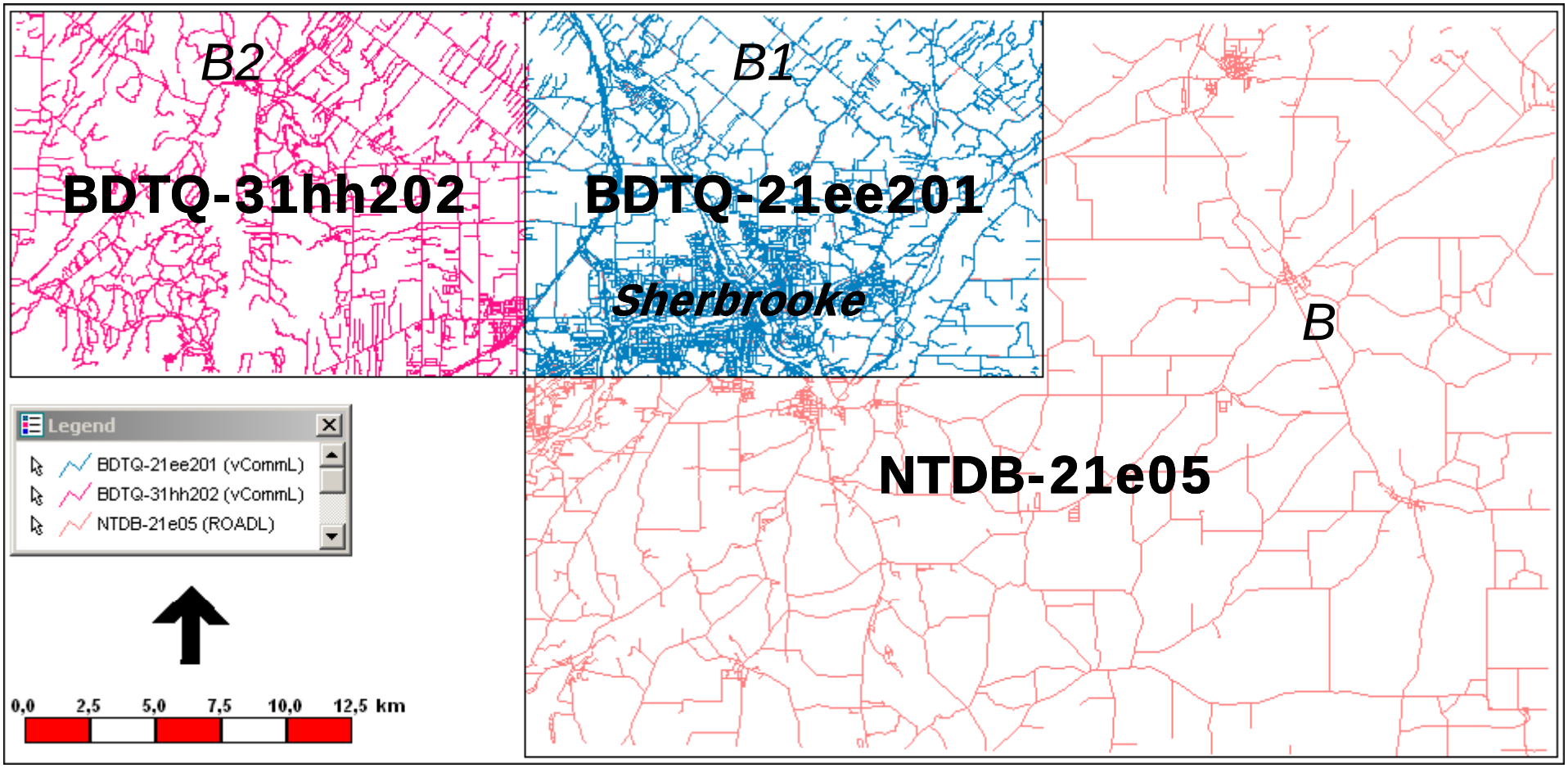
Data Quality Information		Quality information for the data specified by a data quality scope
Lineage		Info. about the events or source data used in constructing the data
Source		Info. about the source data used in creating the data specified by the scope
Scale Denominator	SD	Denominator of the representative fraction on a source map
Positional Accuracy		Accuracy of the position of features
Absolute External P. A.	AEPA	Closeness of reported coordinate values accepted as being true
Thematic Accuracy		Accuracy of quantitative attributes and the correctness of non-quantitative Attributes and of the classification of features and their relationships
Thematic Classification Correctness	TCC	Comparison of the classes assigned to features or their attributes to a universe of discourse
Quantitative Attribute Accuracy	QAA	Accuracy of quantitative attributes
Non-Quantitative Attribute Accuracy	NQAA	Accuracy of non-quantitative attributes
Temporal Accuracy		Accuracy of the temporal attributes and temporal relationships of features
Accuracy of Time Measurement	ATM	Correctness of the temporal references of an item (reporting of error in time measurement)
Temporal Validity	TV	Validity of data specified by the scope with respect to time

Metadata Model

	Feature Class	Geometric Property	Temporal Property	Quantitative Property	Non-quantitative Property	Classification Property	Metadata Element Type
L	√				√	√	String
DT	√						String
DP	√	√	√	√	√	√	Date
BB	√						Bounding Box
VE	√						Z Interval [meter]
TE	√						Time Interval
[R]-ON	√	√	√	√	√	√	String
RSC		√					String
RSCS		√					String
SD		√					Float
AEPA		√					Float [meter]
TCC	√					√	Percentage
QAA				√			Float
NQAA					√		Enum./Number
ATM			√				Float [time unit]
TV	√	√		√	√	√	Time interval

Example

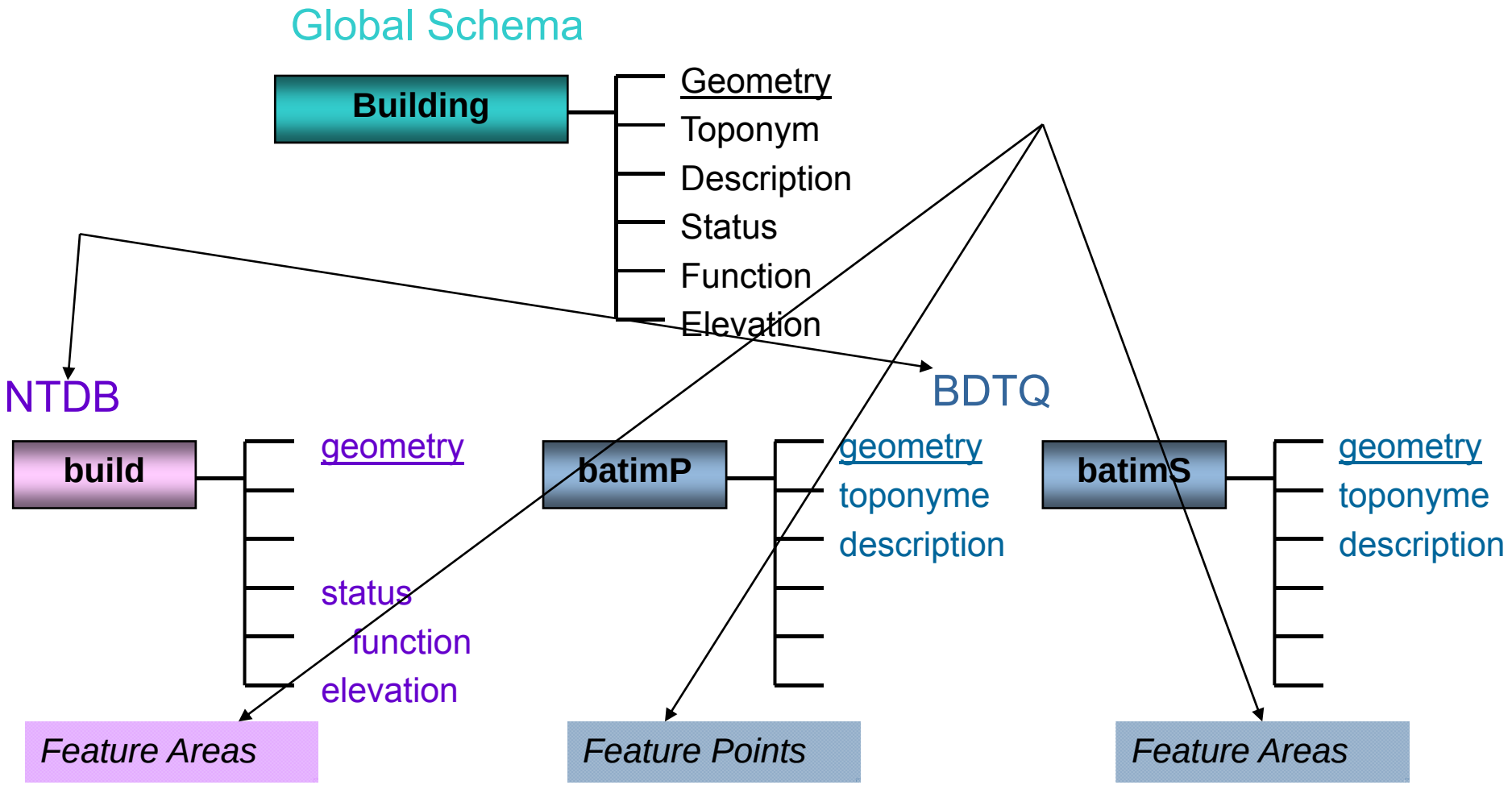
- Three data sources



Acknowledgement : Ministry of Natural Resources Québec

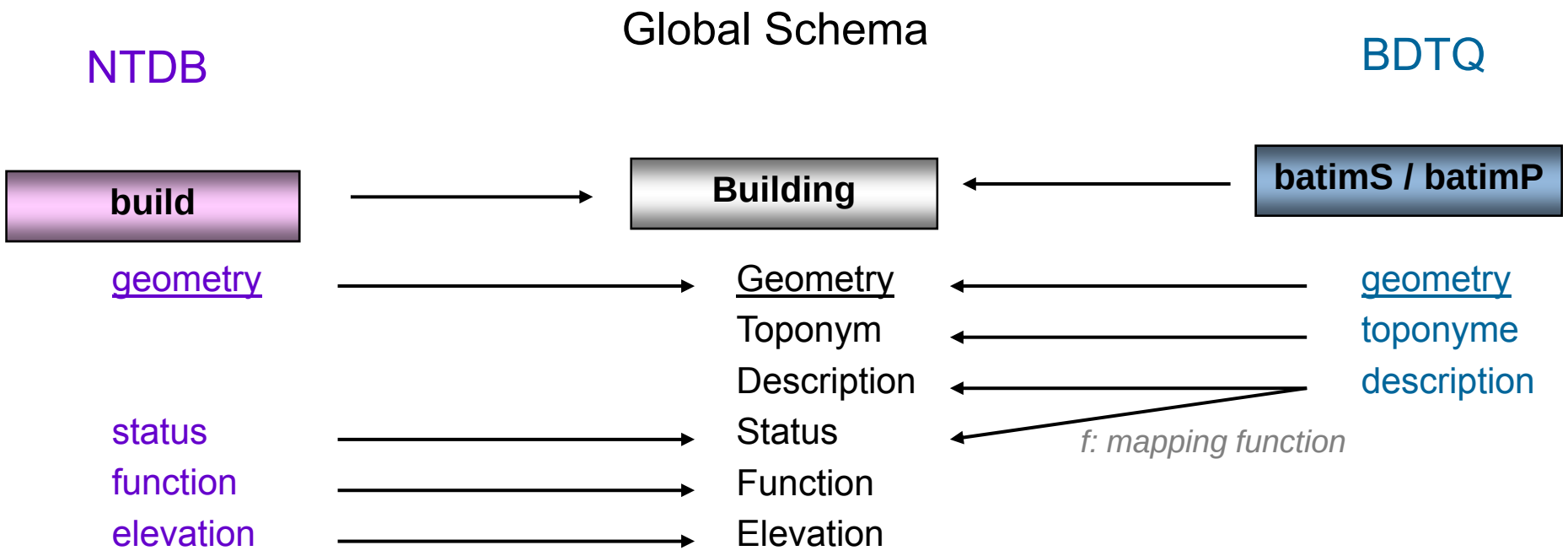
Example

- Data Schemas



Example

- Mappings between Schemas



Example

- Data sources' metadata: BDTQ-21ee201 & 31hh202

	batimP / batimS	geometry	description	toponym
L: Language	French		French	French
DT: Dataset Title	BDTQ-*X*			
DP: Date – Publication	1995	1995	1995	1995
BB: Geographic Bounding Box	B1 / B2			
[Pvd]-ON: Provider – Organization Name	Ministry of Natural Resources - Canada			QTC
RSC: Reference system – Code		WGS84		
RSCS: Reference System – Code Space		WGS		
SD: Scale Denominator		20 000		
AEPA: Absolute External Positional Accuracy		4 meters		
TCC: Thematic Classification Correctness	1%			
QAA: Quantitative Attribute Accuracy				
NQAA: Non-Quantitative Attribute Accuracy				

Example

- Data sources' metadata: NTDB-21e05

	buildid	geometry	status	function	function
L: Language	English		English	English	
DT: Dataset Title	NTDB-*X*				
DP: Date – Publication	1996	1996	1996	1996	1996
BB: Geographic Bounding Box	B				
[Pvd]-ON: Provider – Organization Name	**Topographic Information Center**				
RSC: Reference system – Code		WGS84			
RSCS: Reference System – Code Space		WGS			
SD: Scale Denominator		50 000			
AEPA: Absolute External Positional Accuracy		10 meters			
TCC: Thematic Classification Correctness	5%			5%	
QAA: Quantitative Attribute Accuracy					5 meters
NQAA: Non-Quantitative Attribute Accuracy					

Mapping Rules

BDTQ-21ee201

	L	DP	BB	AEPA	NQAA	DT	*[Pvd]*-ON	RSC	RSCS	SD	TCC	QAA
Building ← batimS / batimP	French	1995	B1			BD…	MNR				1%	
Geometry ← geometry		1995		4*m*			MNR	WGS84	WGS	20000		
Toponym ← toponyme	French	1995					QTC					
Description ← description	French	1995			Medium		MNR					
Status ← f (description)	French	1995			Low		MNR					

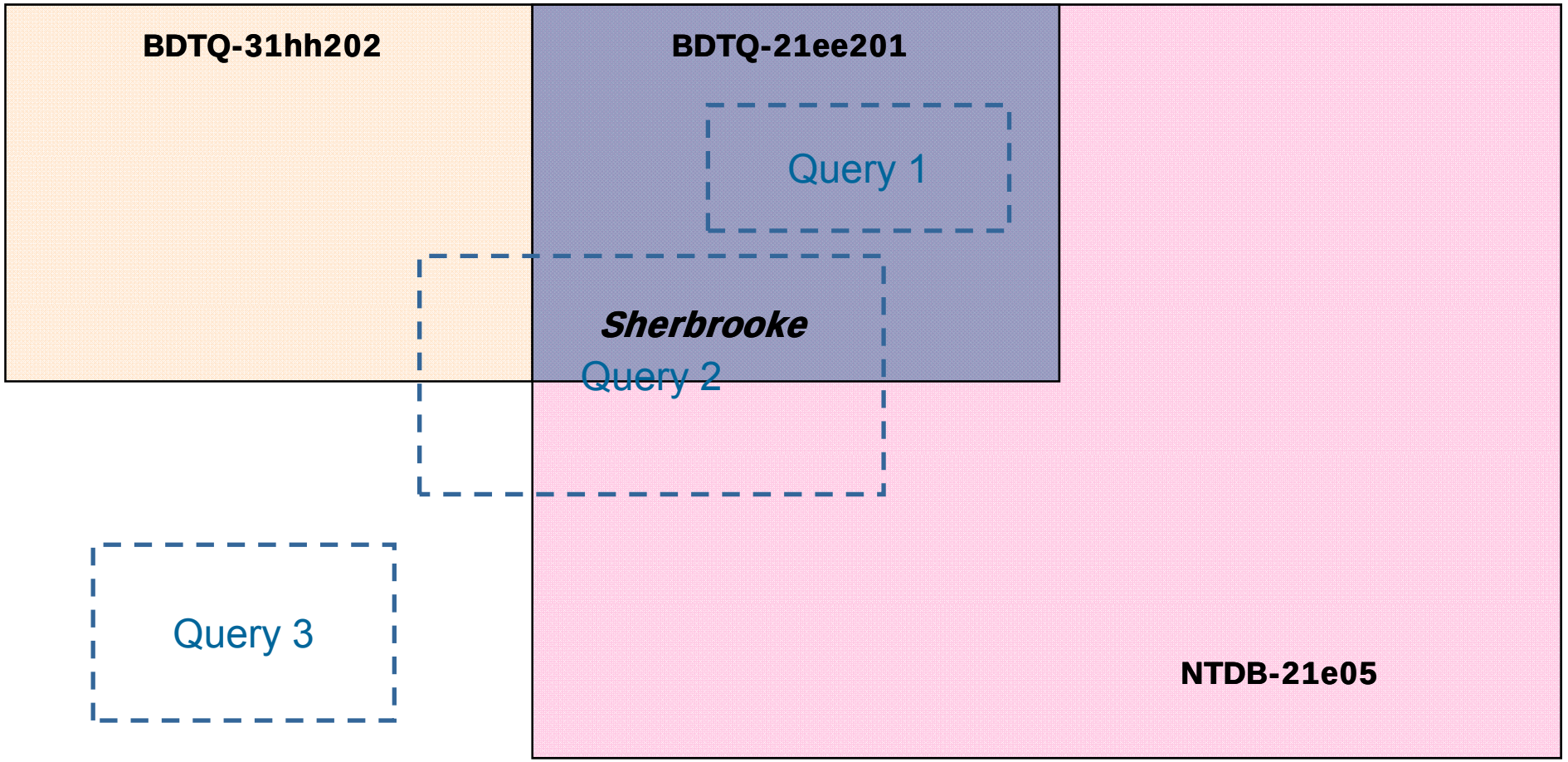
NTDB-21e05

Building ← build	English	1996	B			BD…	TIC				1%	
Geometry ← geometry		1996		10*m*			TIC	WGS84	WGS	20000		
Function ← function	English	1996					TIC					
Elevation ← elevation		1996					TIC					
Status ← status	English	1996			High		TIC					

Rules for BDTQ-31hh202 are similar to those for BDTQ-21ee201 except for the bounding box.

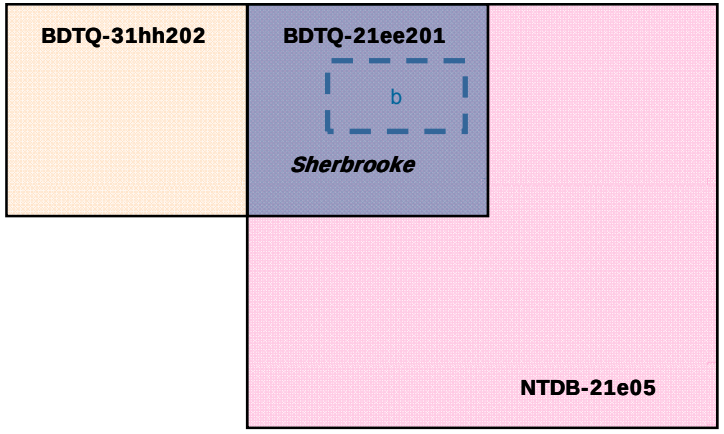
Query Processing

- Select all information about buildings



Example Query

- Query Q1:

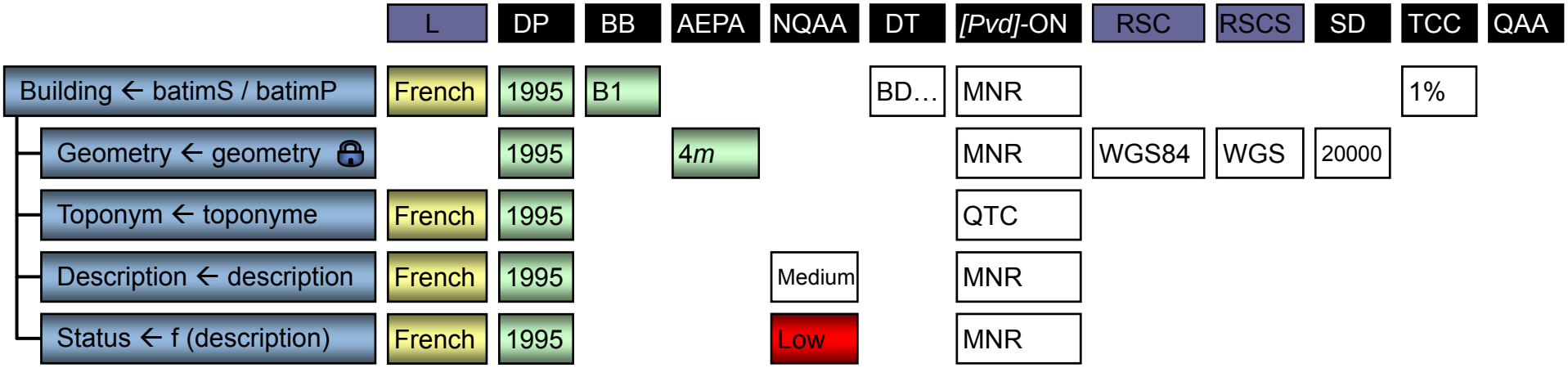


Select all information about buildings,
such that:

- data cover region b
- data published during the last 15 years
- geometric accuracy is at most 5 meters
- operation status of buildings is of **high accuracy**
- Data language is English

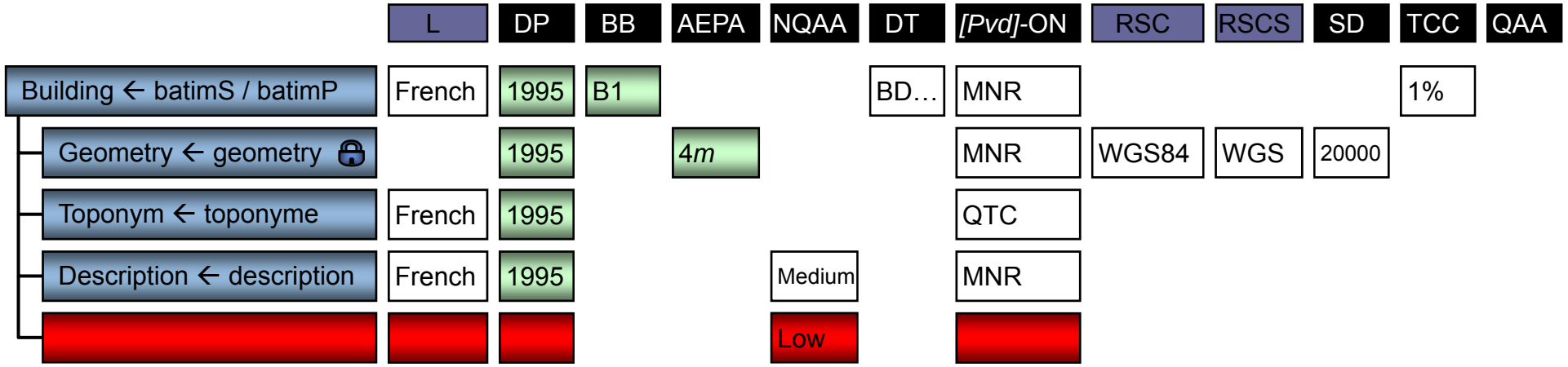
Query Rewriting

BDTQ-21ee201



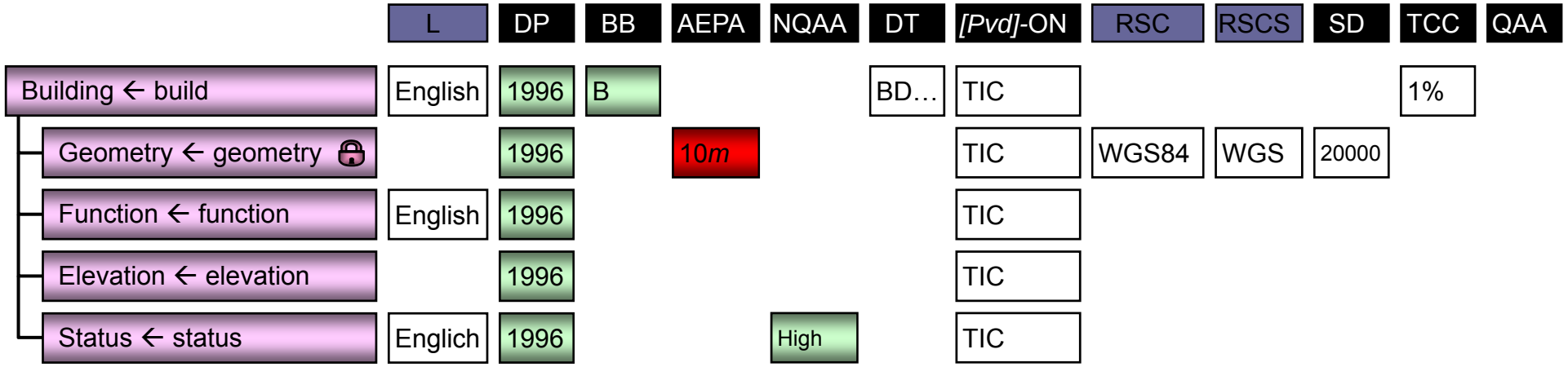
Query Rewriting

BDTQ-21ee201



Query Rewriting

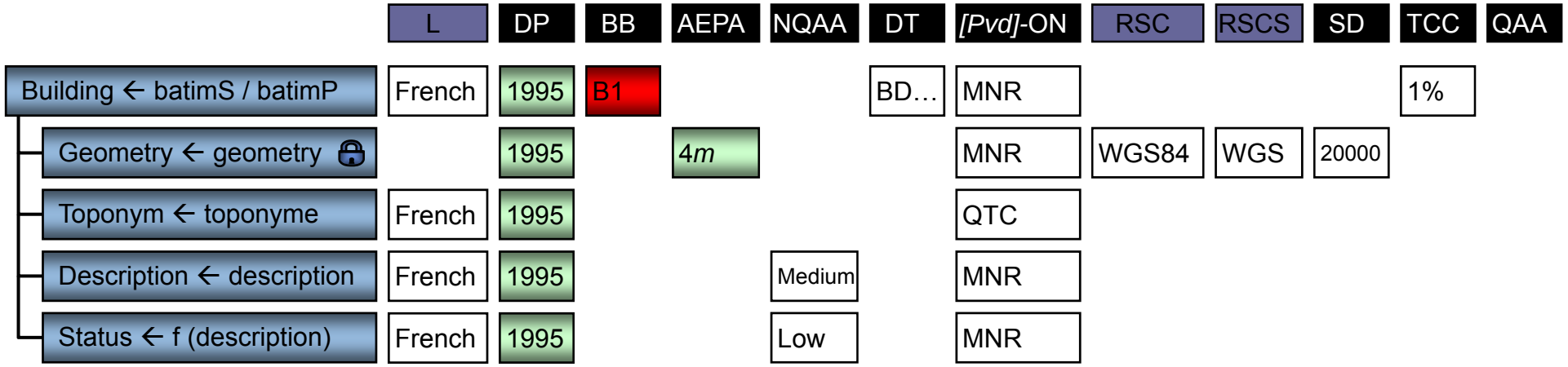
NTDB-21e05



Although the geometric property is not accurate enough, we need to keep it in order to join information with other data sources.

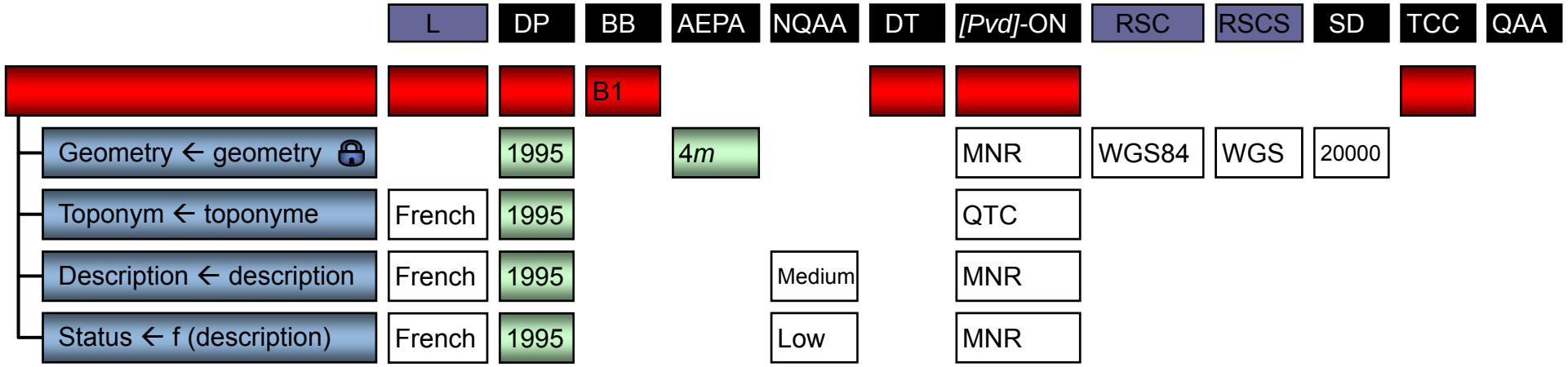
Query Rewriting

BDTQ-31hh202



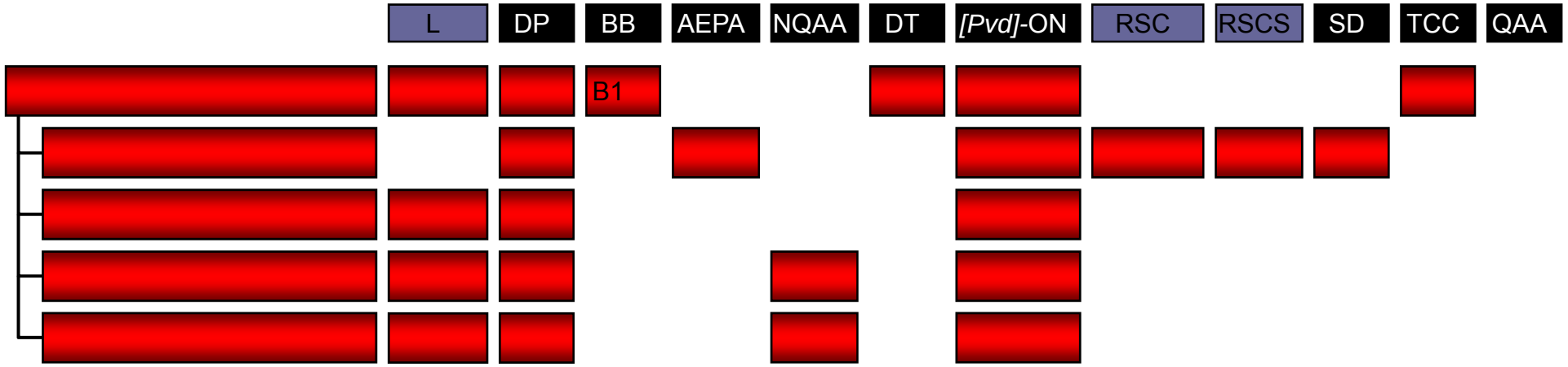
Query Rewriting

BDTQ-31hh202

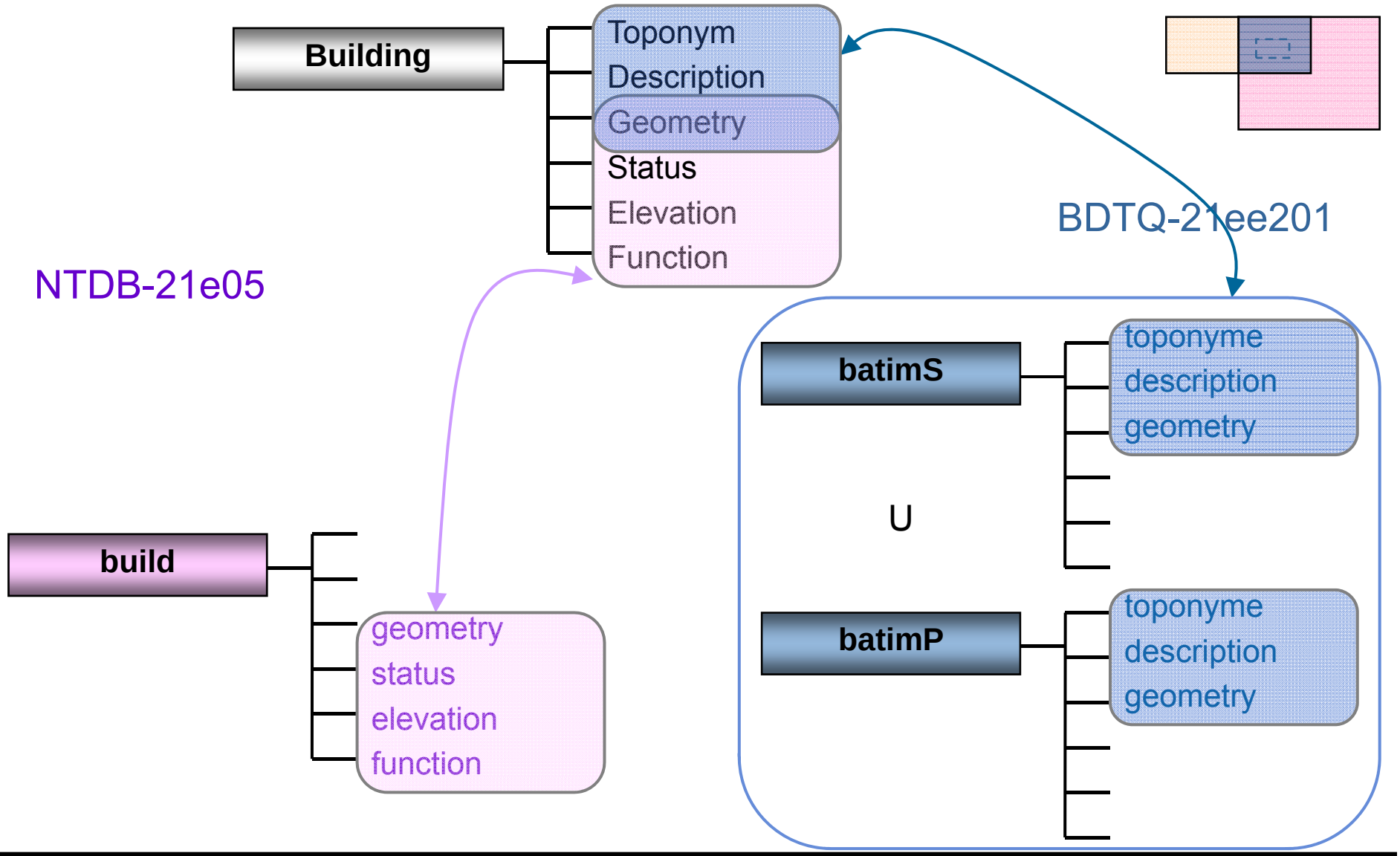


Query Rewriting

BDTQ-31hh202

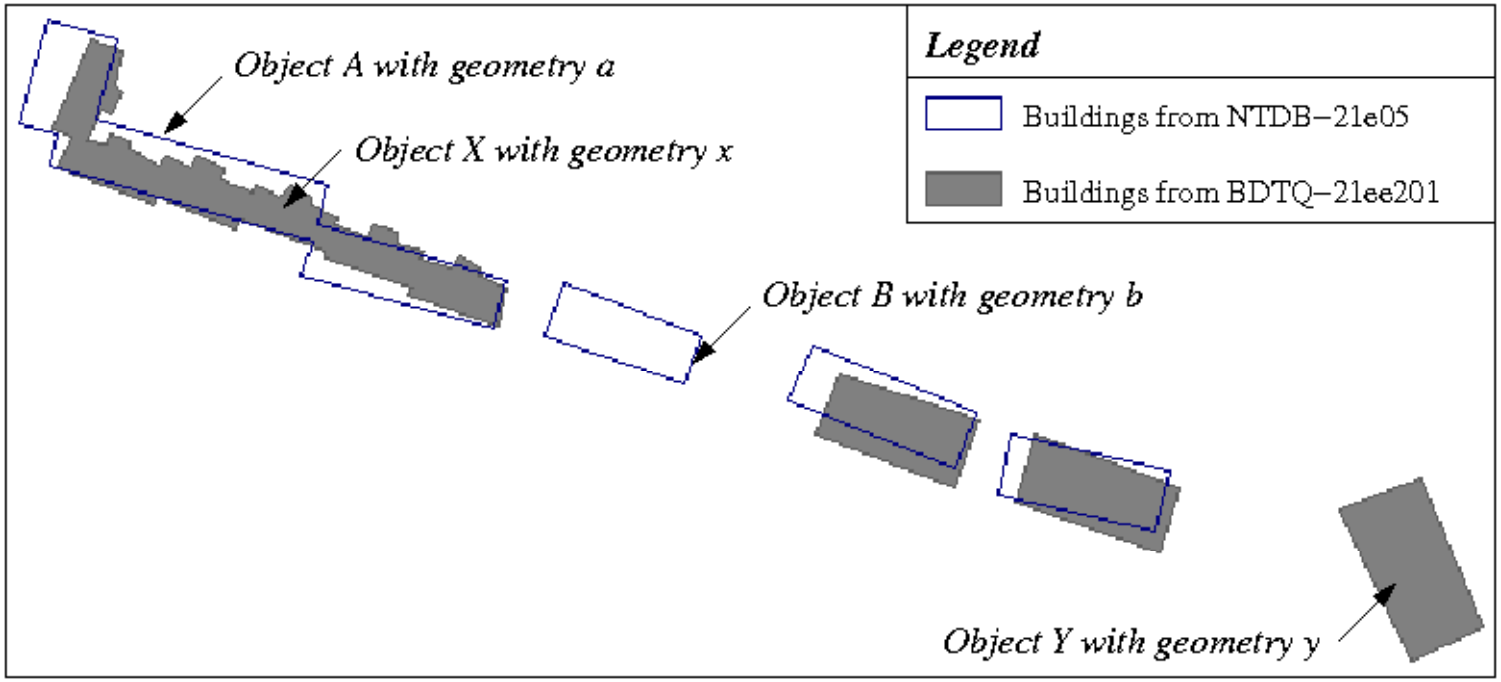


Query Rewriting



SP Query Rewriting

- Object Fusion: A costly process



Object A with geometry a

```
<buildID>
  <id>151</id>
  <geom>a</geom>
  <function>hospital</function>
  <elevation>45</elevation>
</buildID>
```

⋈

Object X with geometry x

```
<batimS>
  <id>24</id>
  <geom>x</geom>
  <toponym>CHUS</toponym>
  <description>...</description>
</buildID>
```

→

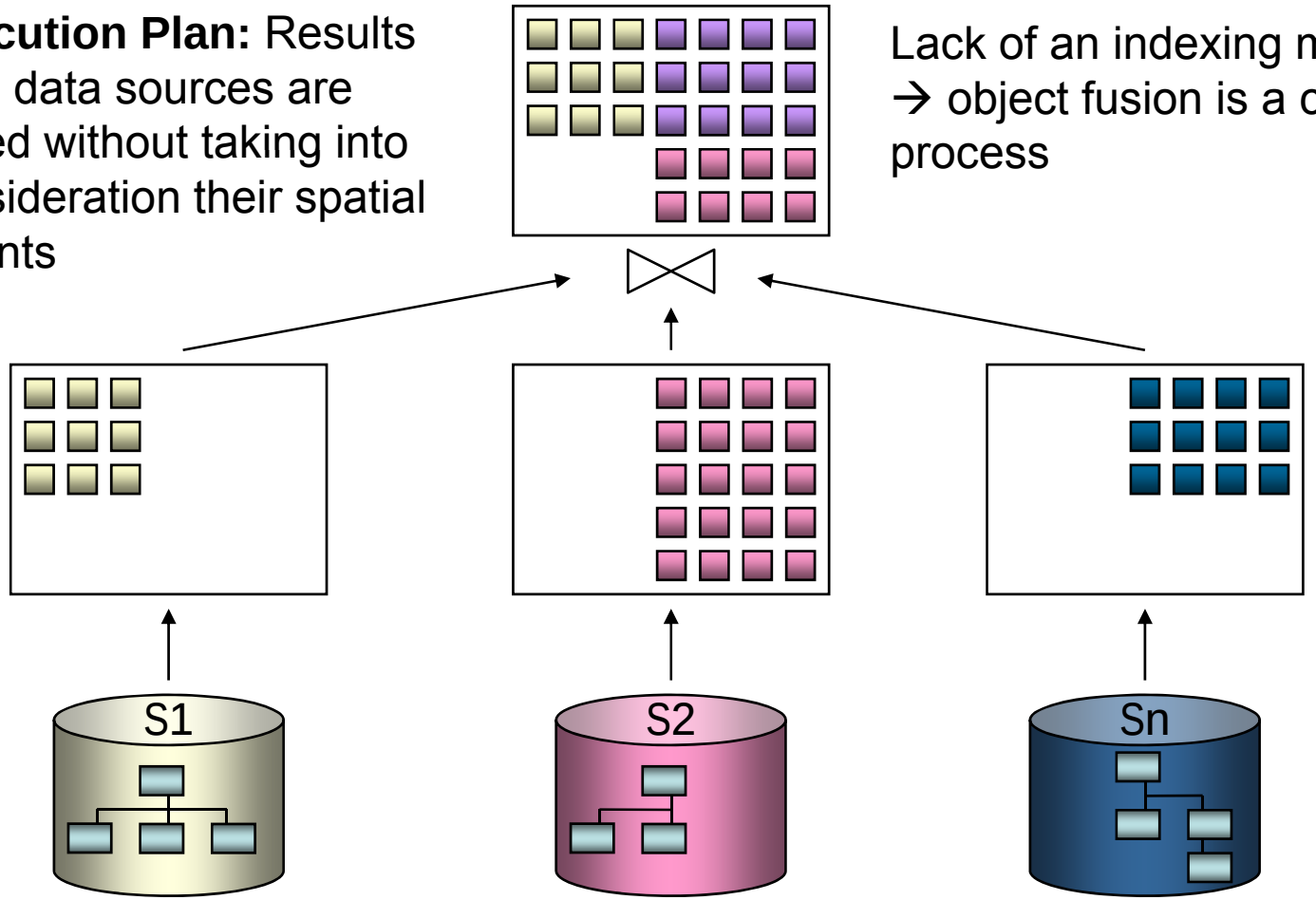
Fusion (join) of A and X

```
<Building>
  <Geom>a</Geom>
  <Function>hospital</Function>
  <Elevation>45</Elevation>
  <Toponym>CHUS</Toponym>
  <Description>...</Description>
</Building>
```

SP Query Rewriting

- Classical mediation approaches

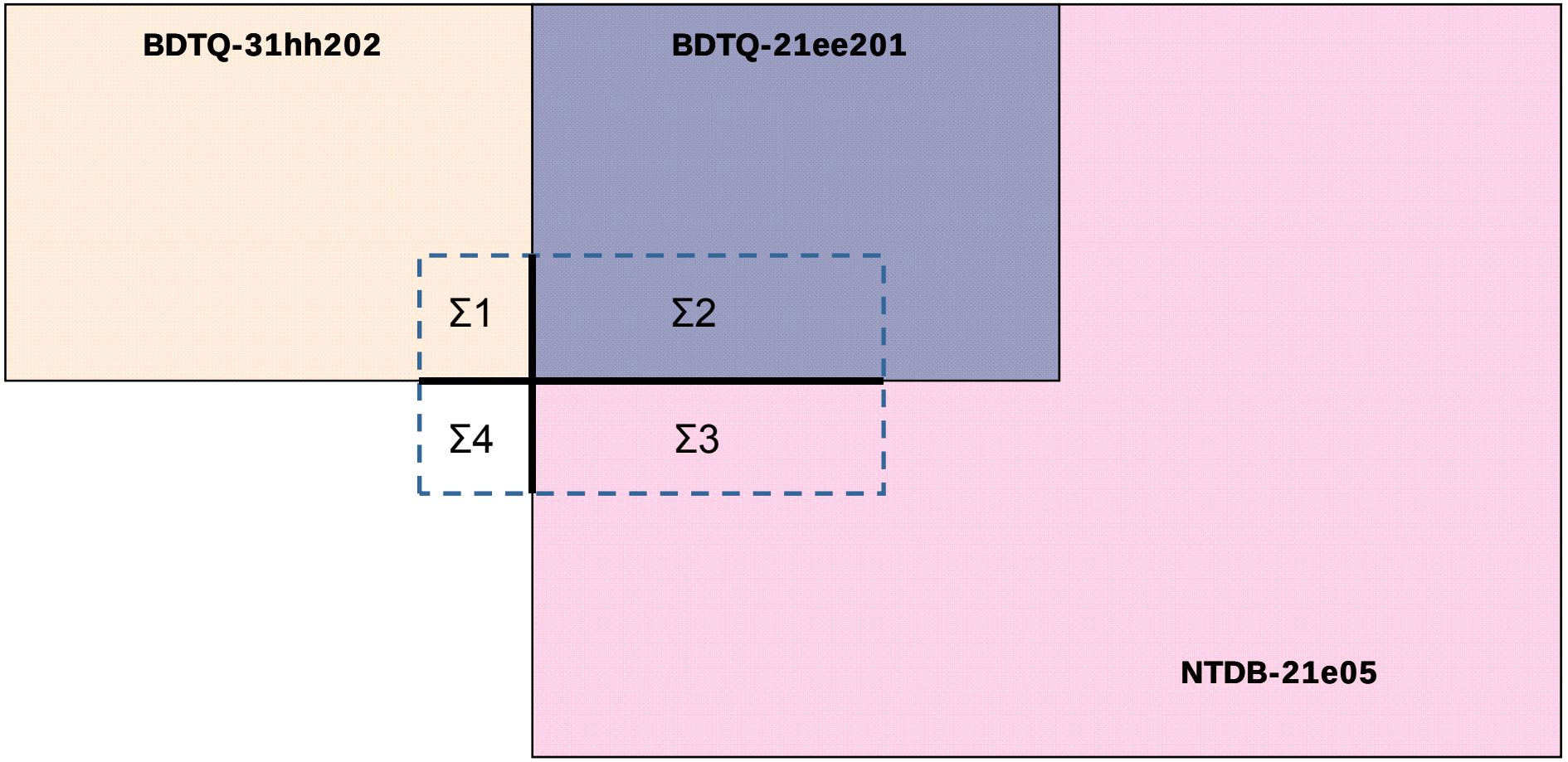
Execution Plan: Results from data sources are joined without taking into consideration their spatial extents



Lack of an indexing mechanism → object fusion is a costly process

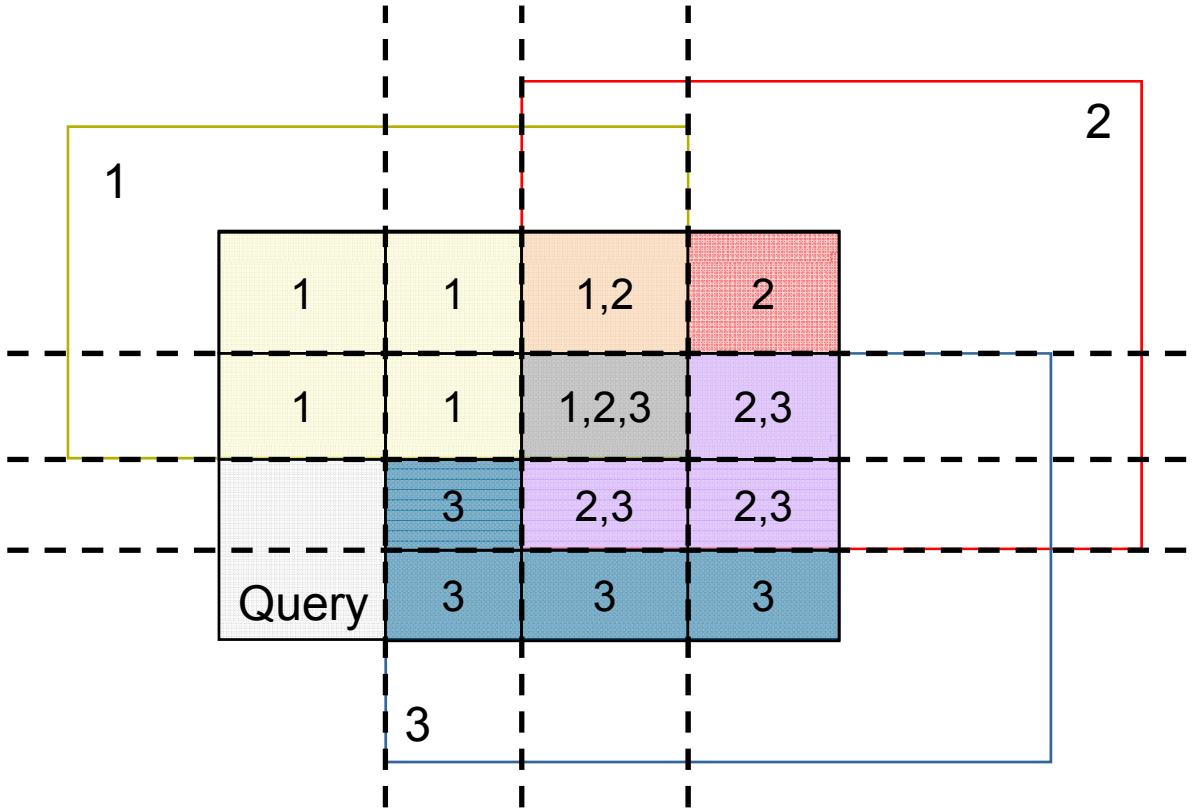
SP Query Rewriting

• Space Partitioning: The query bounding box is divided divided into sectors



SP Query Rewriting

• Space Partitioning: A query execution plan per sector: *Sectoral Execution Plan*



Benefits of Space Partitioning

- Reduce the number of useless join operations (join of objects from different regions)
- Increase the number of simple operations (data retrieval), which can be parallelized
- Decrease the number of complex operations (spatial join operations)

Experimental Results

Size of the query's bounding box

Number of data sources

Number of generated Sectors

Execution time (s) with space partitioning

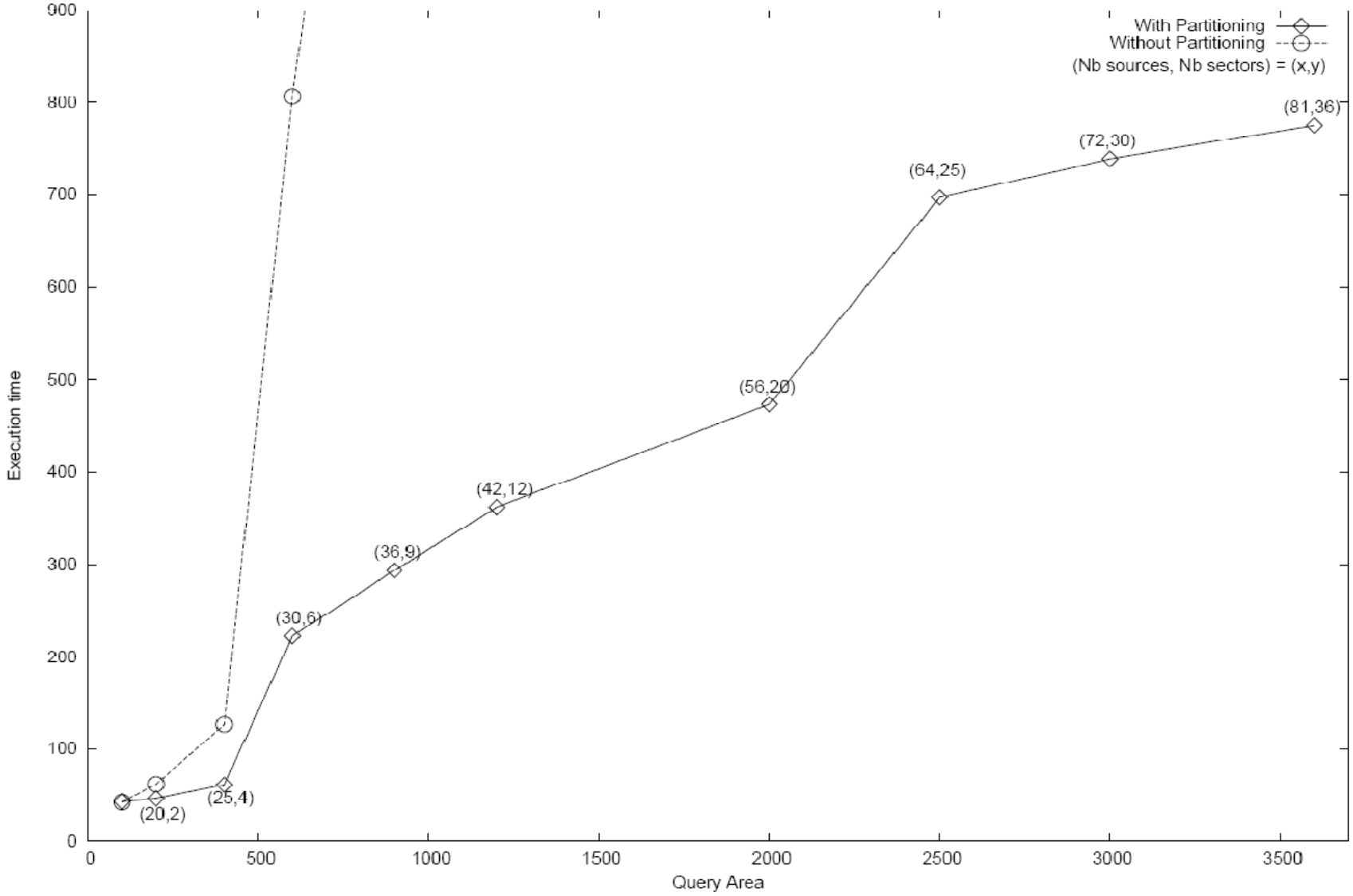
Execution time (s) without space partitioning

S_Q (%)	n	m	Spl Algo	S-A Algo	Gain
2	6	2	10	12	12,37
4	9	4	16	29	43,80
6	12	6	52	180	70,88
9	16	9	51	390	86,67
12	20	12	55	408	86,53
16	25	16	77	654	88,14
20	30	20	100	1579	93,63
25	36	25	160	3091	94,81
30	42	30	170	5029	96,61
36	49	36	186		

Data:

SEQUOIA 2000 regional benchmark point Data

Experimental Results



Conclusion and Future Work

- What has been done so far
  - Quality-mediator for geographic data sources
  - A space partitioning technique for query optimization
- Future Work
  - Generalize the approach for more metadata elements,
  - Develop a Web services oriented architecture.