A Measurement Set Data Model

BASP, Villars **Sept. 8 2011**

F. Viallefond.

Laboratoire d'Étude du Rayonnement et de la Matière en Astrophysique

Outline

1. What is a Dataset?

2. Motivations, experiences, evolutions

3. What is a Data Model?

4. Formalization

   - Categories, Functors

5. Two examples showing this formalism at work

   - Physical quantities

   - Physical measurements in context, measurement set.

6. Conclusions

Dataset

A dataset contains every things needed to make the raw observational data scientifically useful (science archive, off-line data reduction and analysis)

1. Correlator data
   (99% of the amount of data, simple compact structure)

2. Metadata

   - target space parameters (astronomical positions: directions, spectral (line) frequencies, ...)
   - instrumental configurations used
   - experimental procedures (observing modes)
   - instrumental parameters used (antenna positions, pointing model used, ...)
   - fine tuning data (quasi real time calibration results)

3. Auxiliary data

   - encoder readouts (dish) / complex gains (aperture arrays), etc...
   - experimental context (antenna-based monitoring points: temperature, pressure, ... atmospheric radiometric data, ...)

Motivations to have a Data Model

A measurement set is a set of concrete concepts at different levels,

a) words, *e.g.* physical quantities, measures (Universal Concepts),

b) compositions of words defining relations (Domain Specific Concepts).

Common language & understanding of concepts (interoperability).

a) expressiveness

b) robustness (type-safe)

c) efficiency (static typing, high performance calculi, ...)

(architecture: structure, factorization, localization, slicing, ... i.e. geometry),

The model must be as rich as needed within a context evolving towards more and more automated processing (data volume, instrumental complexity, ...)

From acquired Experiences to required Evolutions

Experiences:
The radioastronomy has accumulated knowledges and experiences for many years

Evolution from data formats to DMs
major step in 1995/2000 with MS (ref. Kemball et al.)

Broader usages:
a) for persistence (archives),
b) for off-line data processing (software packages, pipelined processing, ...)
c) for on-line data acquisition (near real time telescope calibration, quick look, ...)

NB: *transporting data is time consuming* $\rightarrow$ *data flows must be well thought*

Instrumental evolution: begs for DM evolutions.
Example: aperture arrays like EMBRACE (proto for SKA)

Facts: the mathematicians:
a) have developped all the abstract constructs useful to us
b) give a methodology to define data models & theories *(ref. theory of categories)*
NB:
a) formalism used in fundamental computer science.
b) matchs well with generic programming techniques.

EMBRACE

Storage capacity to record N beamlets (GBytes or Tbytes)

beamlets	data flow	10 min	1 hour	5 hours	10 hours
62	93 MB/s	54.2	325.5	1.6	3.2
124	185 MB/s	108.5	651	3.2	6.4
186	278 MB/s	162.8	976.4	4.8	9.6
248	370 MB/s	217	1.3	6.4	12.8

What is a data model?

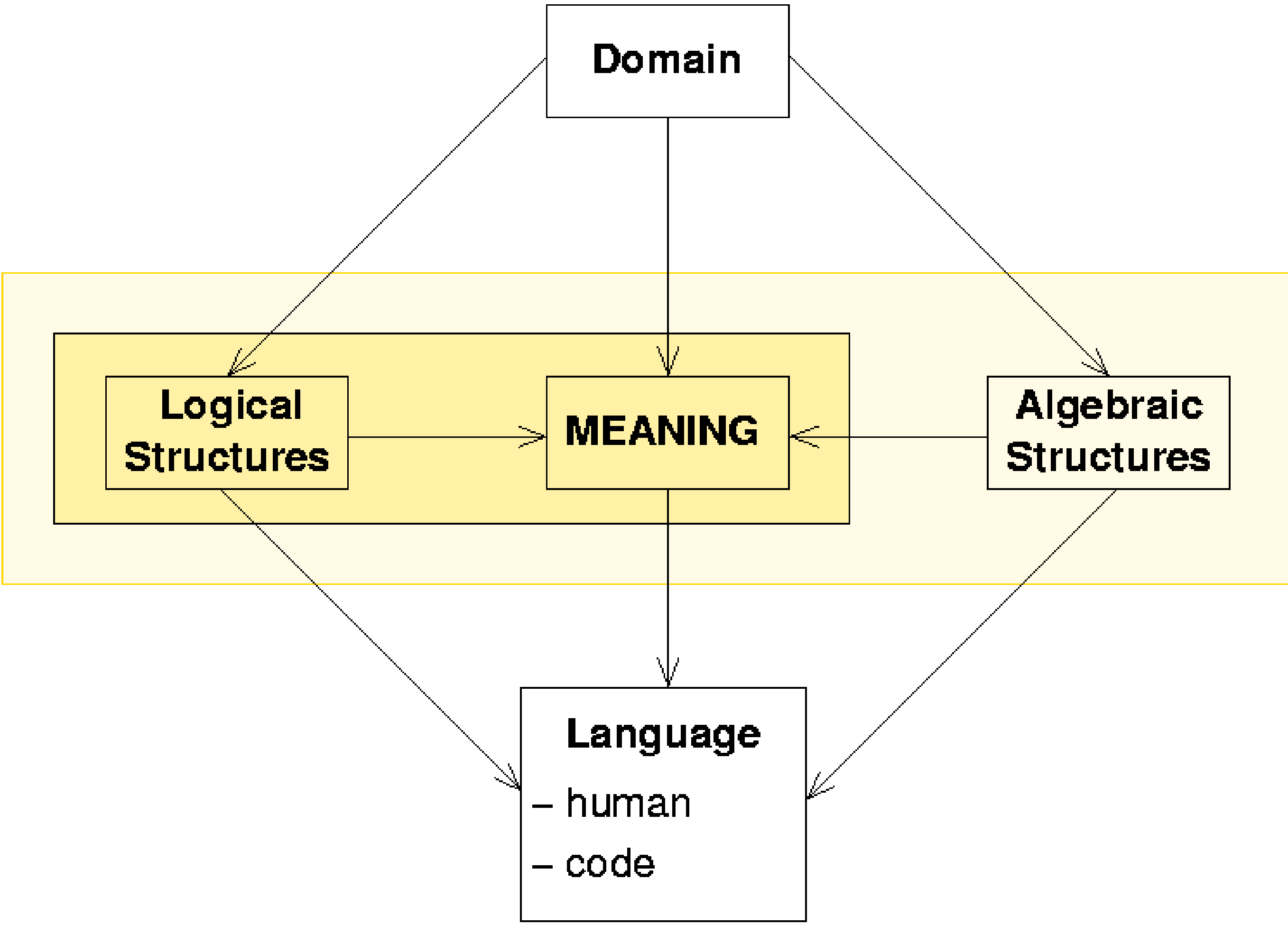
A model is the composition of
a structure (mathematical logic) with algebra.

Example: the relational data model.

- The semantic is captured through constraints.
- The structure gives the meaning of things in a formal language.

Datasets must conform to a model

4 commutable triangles



To use a language for representing measurements

Examples of words *(physical quantities)*:

- Length, Area, Angle, Solid angle, Aperture efficiency, Rotation measure
- Speed
- Angular rate
- Noise equivalent power
- FluxDensity *(Jy which is not SI...)*
- ...

Note that:

1. All these have units.
2. Dimensioned, dimensionless and mixed case units!
3. They may have units which uses powers of rational numbers!
4. Physical expressions are composition of such words

Measurements in context

We assign domain specific meaning to words:

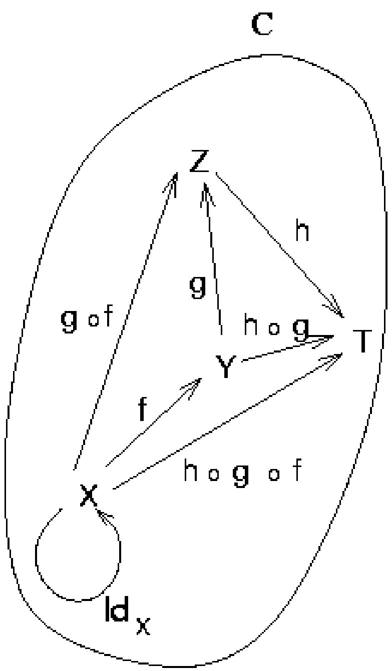
- Station
- Antenna
- Spectral window
- Feed
- Configuration description
- ...

Meta-model $\rightarrow$ meta-model instance $\leftarrow$ a DSL

Formalization

- **Category**
- **Functor**
- *Natural transform*
- *Product and coproduct:*
  example of diagrams, a cone (projections) and a cocone (inductions)
- *Direct limit*
- *Monoids. 2-categories, ...*
- *Sketches, Models and Theories*

Category
C



Collection of objects:

X, Y, Z, T

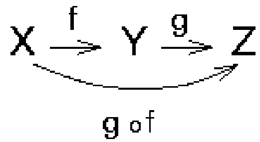
Morphisms of objects:

f, g, h

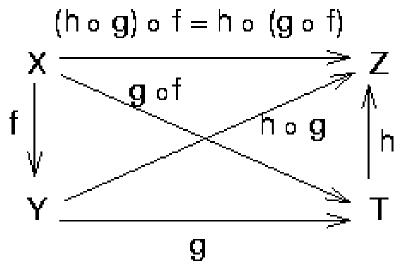
- Identity:

  $\forall \; X \in \mathbf{C} \; \exists \; Id_X \in \mathbf{C}$

- Transitive composition:

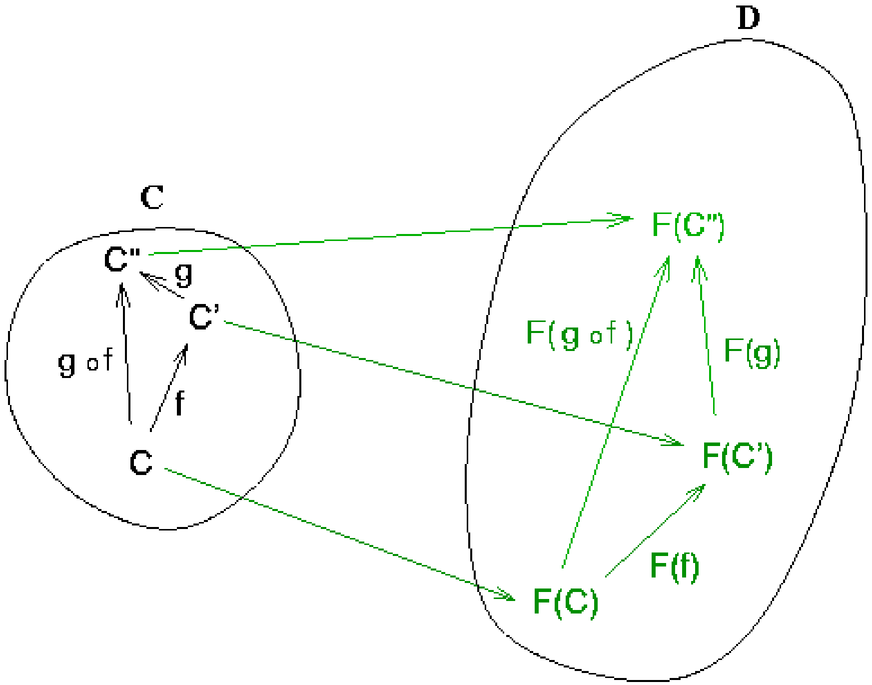


- Associativity:



Functor

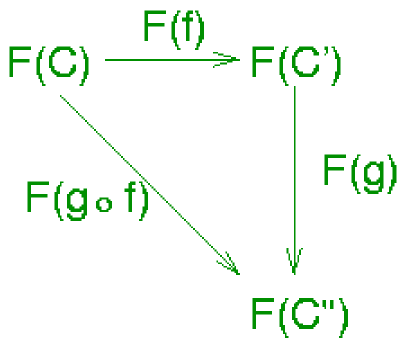
$F: \mathbf{C} \longrightarrow \mathbf{D}$



Two categories $\mathbf{C}$ and $\mathbf{D}$

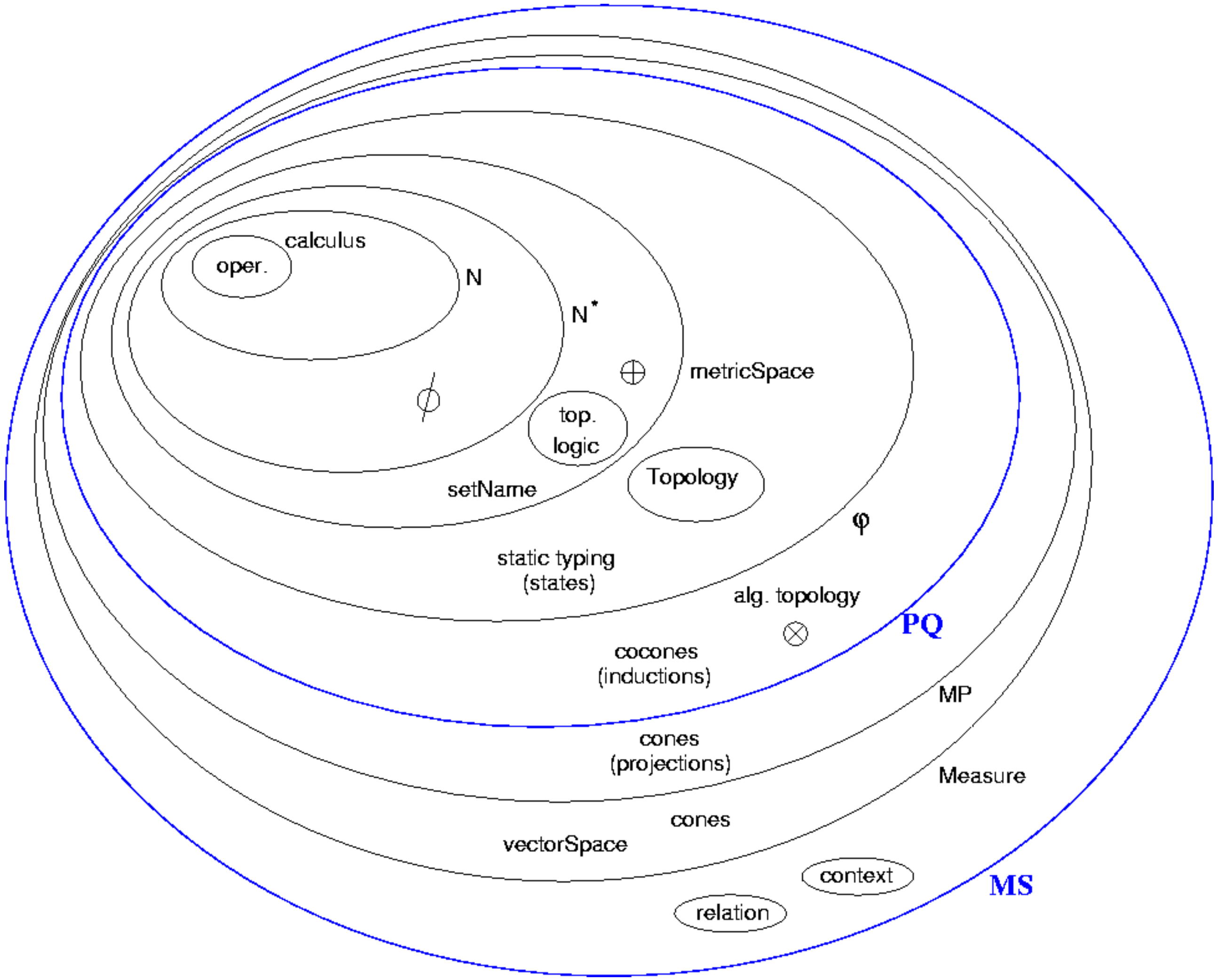
The morphism $F: \mathbf{C} \longrightarrow \mathbf{D}$ is a functor if:

- $\forall\ C \in \mathbf{C}\ \exists\ F(C) \in \mathbf{D}$
- $F(Id_C) = Id_{F(C)}$
- and the diagram



is commutative

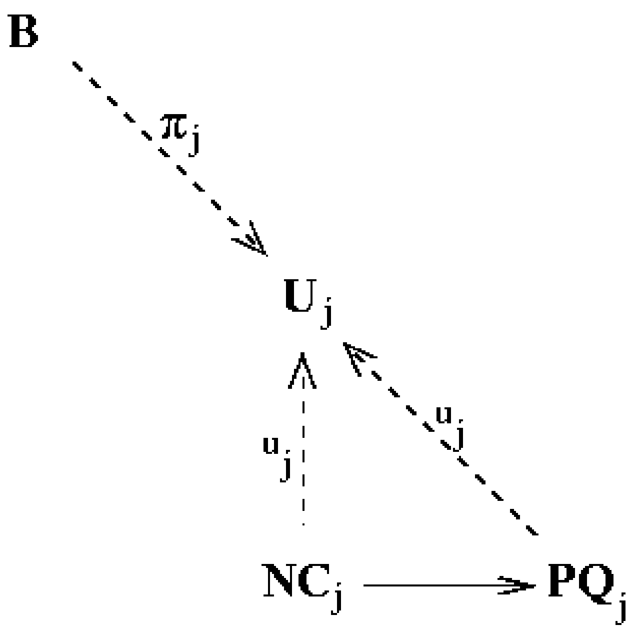
Two examples at work



Topological space for PQ

Let $\mathbf{U}_j$ a unit element along the axis j in a vector space **B** an object $\in$ **Vect**

Let $\mathbf{NC}_j \in \mathbf{R}$ the dimension unit along the axis j of a point in that space



Examples with j=0

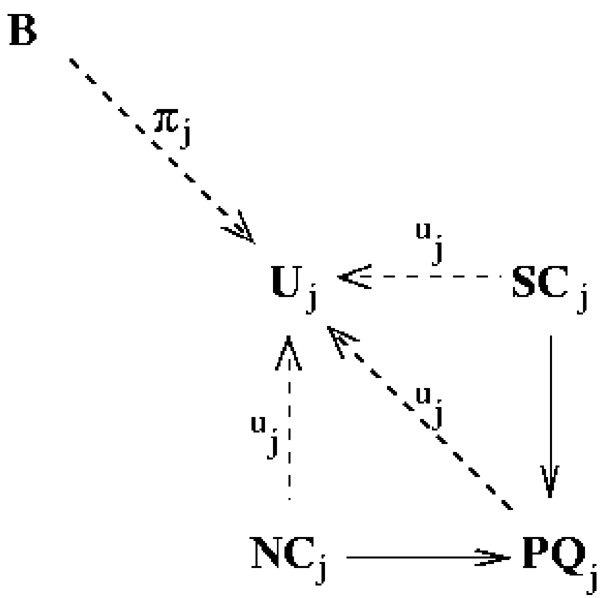
Phys quan. k	$\mathbf{NC}_{0,k}$
Length	*1*
Area	*2*
SpatialFrequency	*−1*

Topological space for PQ

Let $\mathbf{U}_j$ a unit element along the axis j in a vector space **B** an object $\in$ **Vect**

Let $\mathbf{NC}_j \in \mathbf{R}$ the dimension unit along the axis j of a point in that space

Let $\mathbf{SC}_j \in \mathbf{R}$ the dimension unit ratio along the axis j of a point in that space



Examples with j=0

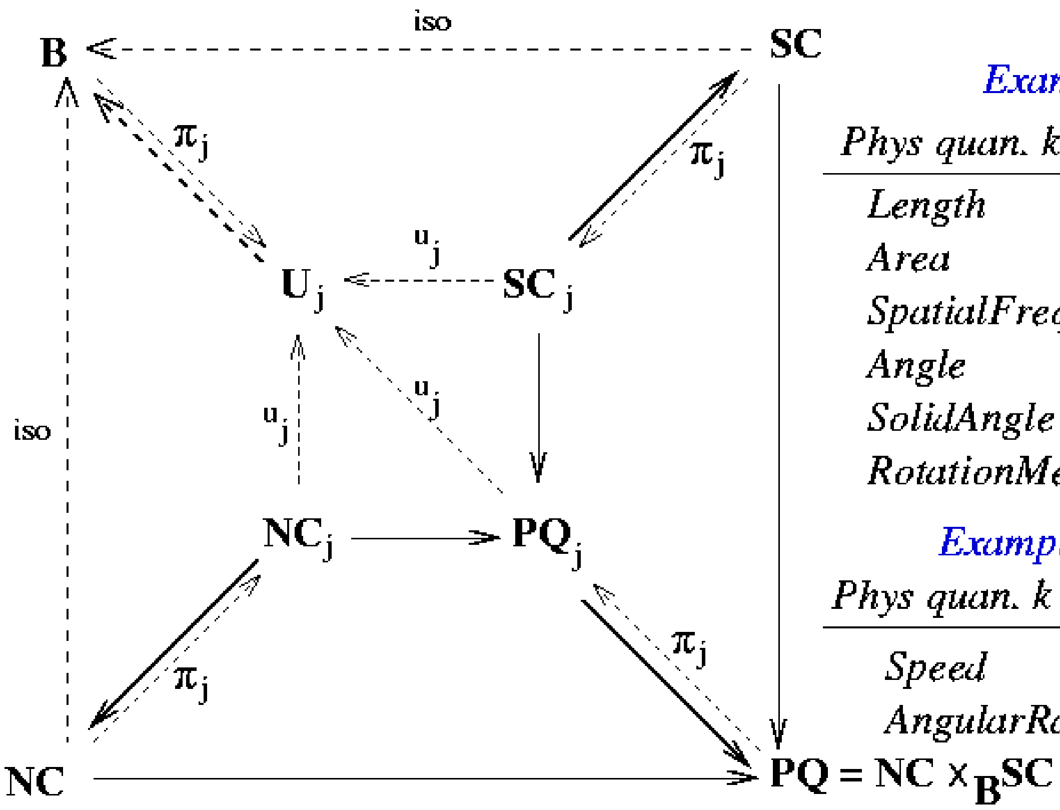
Phys quan. k	$\mathbf{NC}_{0,k}$	$\mathbf{SC}_{0,k}$
Length	*1*	*0*
Area	*2*	*0*
SpatialFrequency	*−1*	*0*
Angle	*0*	*1*
SolidAngle	*0*	*2*
RotationMeasure	*−2*	*1*

Topological space for PQ

Let $\mathbf{U}_j$ a unit element along the axis j in a vector space **B** an object $\in$ **Vect**

Let $\mathbf{NC}_j \in \mathbf{R}$ the dimension unit along the axis j of a point in that space

Let $\mathbf{SC}_j \in \mathbf{R}$ the dimension unit ratio along the axis j of a point in that space



Examples with j=0

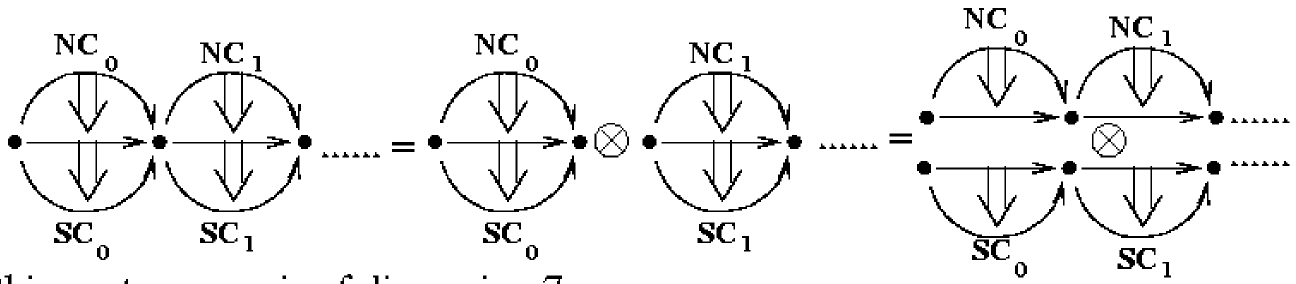
Phys quan. k	$\mathbf{NC}_{0,k}$	$\mathbf{SC}_{0,k}$
Length	*1*	*0*
Area	*2*	*0*
SpatialFrequency	*−1*	*0*
Angle	*0*	*1*
SolidAngle	*0*	*2*
RotationMeasure	*−2*	*1*

Examples with j=0 and 2

Phys quan. k	$\mathbf{NC}_{0,k}$	$\mathbf{SC}_{0,k}$	$\mathbf{NC}_{2,k}$
Speed	*1*	*0*	*−1*
AngularRate	*0*	*1*	*−1*

Topological space for PQ

The topology of PQ is a 2–category on a vector space:



this vector space is of dimension 7:

0 Length
1 Mass
2 Time
3 Temperature
4 LuminousIntensity
5 MolarConcentration
6 ElectricCurrent

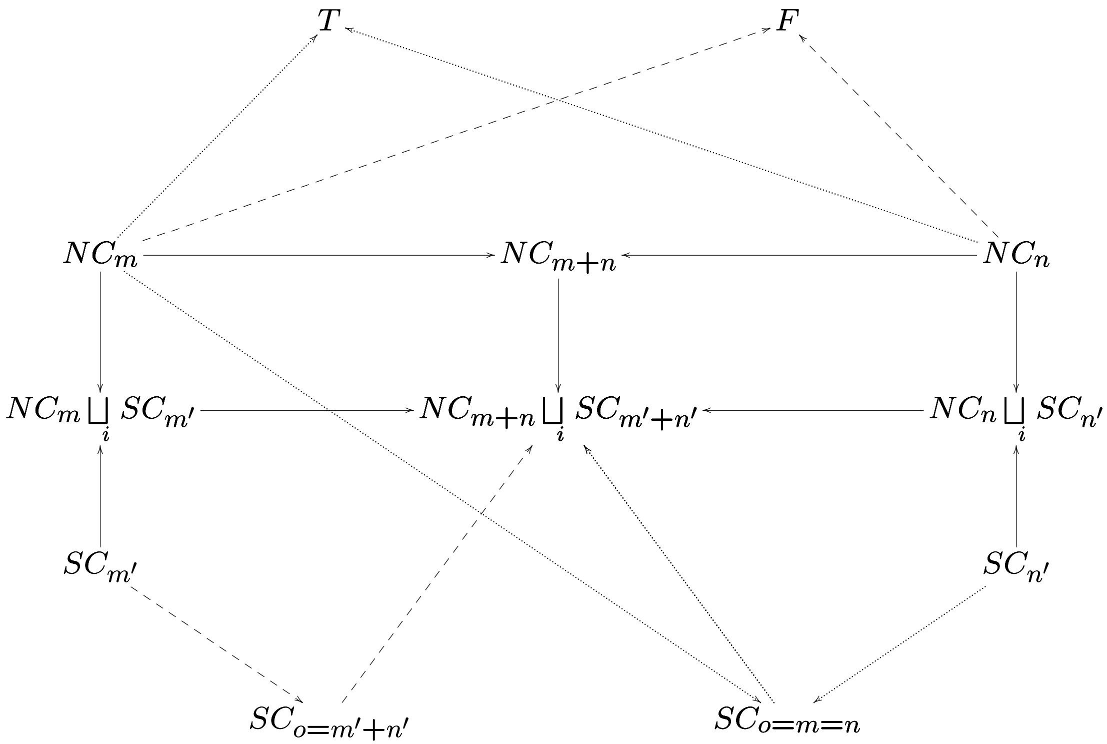
the horizontal 1–cell composition along the foundamental physical unit basis
the vertical 2–cell composition for the dimension,dimensionless property

PQuantity is a monoid for the addition.

PQuantity<X> is a category, a singleton.

From a set of monoids to the category PQ

the algebraic topology

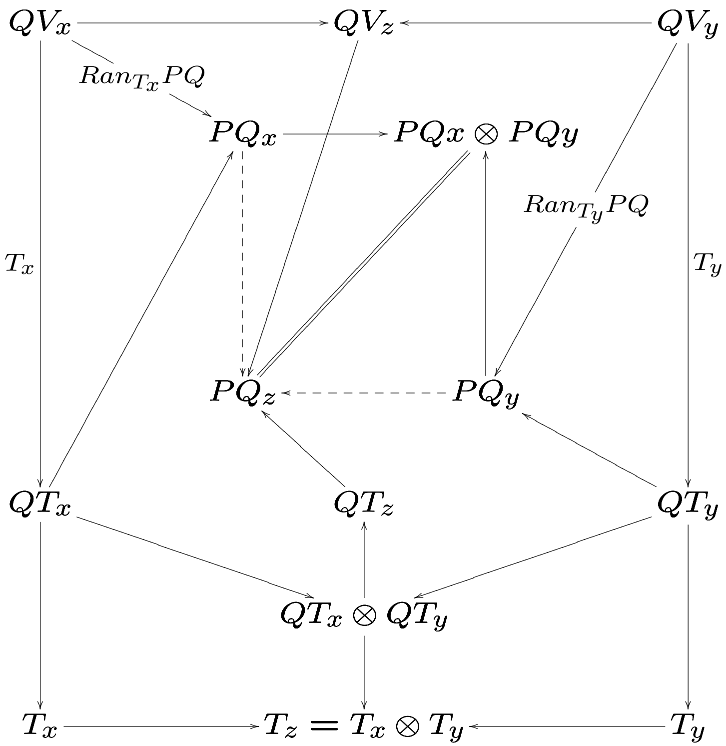


Proposition:

the product of a dimensionless quantity with its inverse is a pure number.

From a set of monoids to the category PQ

Equation of the product: a diagram of PQ



- A linearization on a language (functions basis)
- A coherence constraint (∃ validation)

Logical structure of PQ

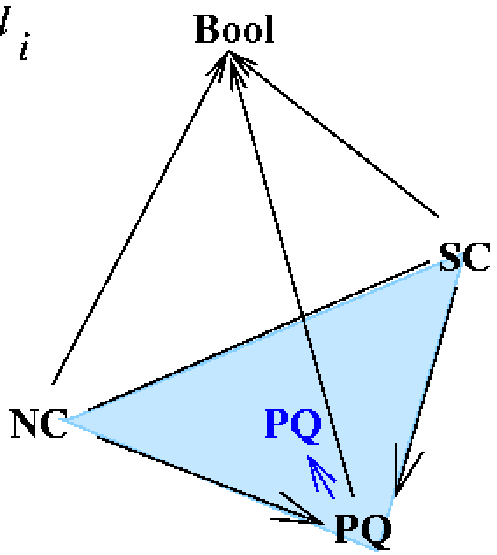
and its boundary

An algebraic type with a closure: $SC_i/SC_i = Id_i$

Identity element: $Id = \bigoplus_{j \in J} Id_j$

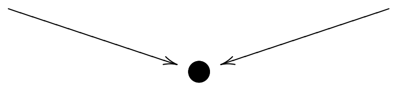
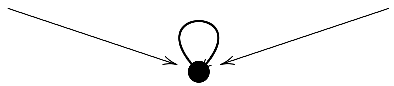
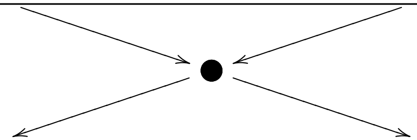
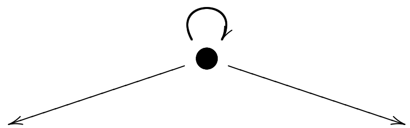
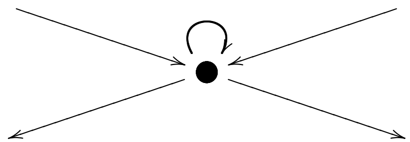
PQ: an endofunctor

Co−end: the pure numbers



Space	Regions in the DSL
2D facette NC,PQ,Bool	sub−category of the dimensionned PQ
2D facette SC,PQ,Bool	sub−category of the dimensionless PQ
3D volume	category PQ: general case

Examples of constructions for the categories PQ and PM

	units		construction	category
•	m		direct	**PQ**
•	rad		inductive	
• •	rad/m		inductive $\oplus$ direct	
• •	$rad \pm \epsilon$		inductive $\oplus$ projective	**PM**
• •	$m \pm \epsilon$		direct $\oplus$ projective	
• • •	$rad/m \pm \epsilon$		inductive $\oplus$ direct $\oplus$ projective	

Conclusions for the Physical Quantities

- PQ is a functor category, a singleton. It is a pure abstraction.
- PQ is the set all the physical expressions
- PQ is an endomorphism
- PQ is a monad PQ(PQ()) = PQ(); $\nVdash \times$ PQ = PQ
- PQ is cartesian closed (*eg* PQuantity is embedded in $\mathbb{R}^*$.)
- PQ$_T$ is a monoid, a constructible functor with polymorphic representation *monomorphism: $Ran_T PQ$ and its dual, $Lan_T PQ$, for polymorphism.*
- PQ$_T$ is a cartesian closed category whose objects are physical quantity states and the morphisms tensor products.
- PQ has inductive cones

Measurement Set Data Model

- Domain specific concepts are build on normalized relations
  (→ keys)

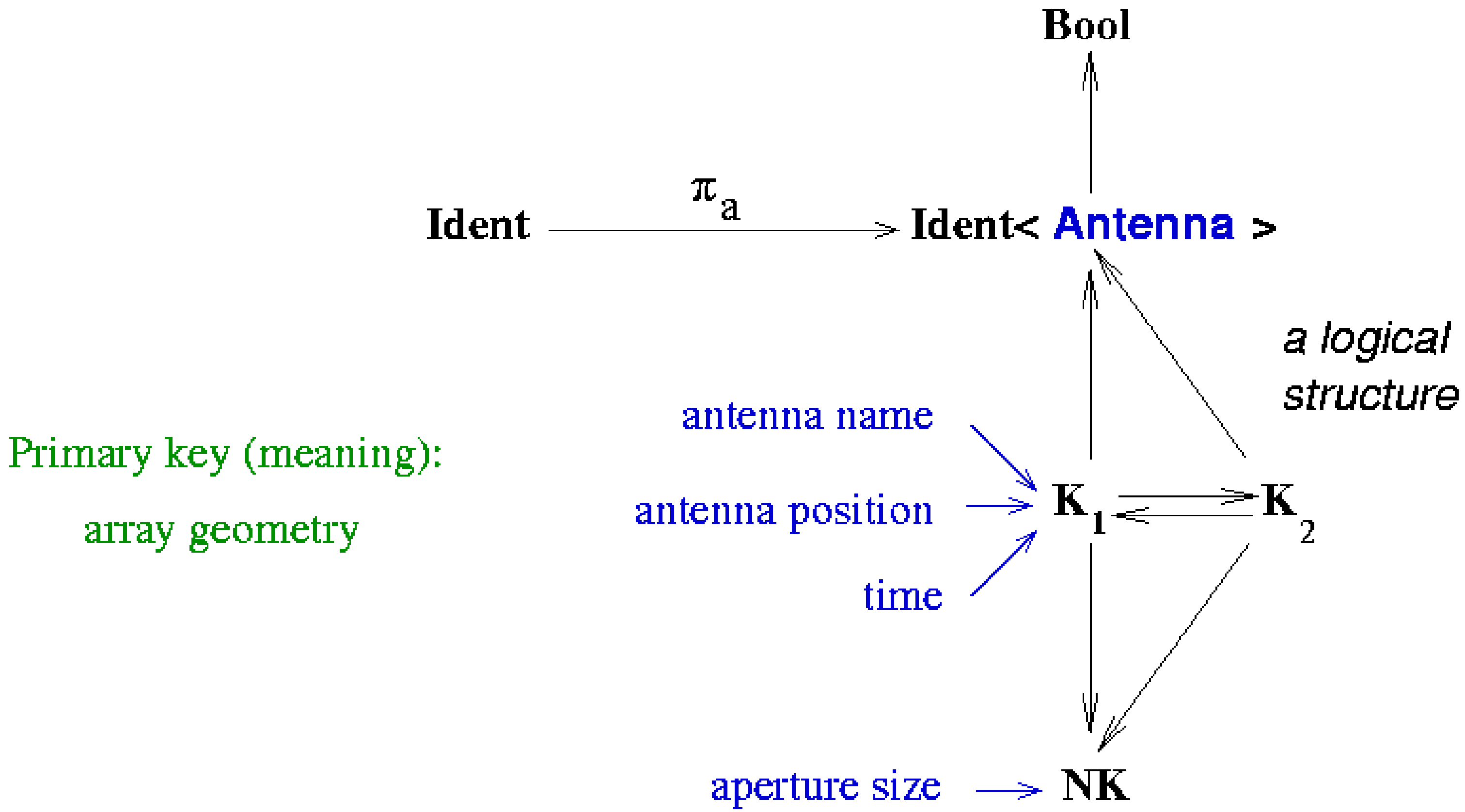
- The measurement set is a set of concepts with relations between them

- Some concepts require objects defined recursively
  (→ model not relational)

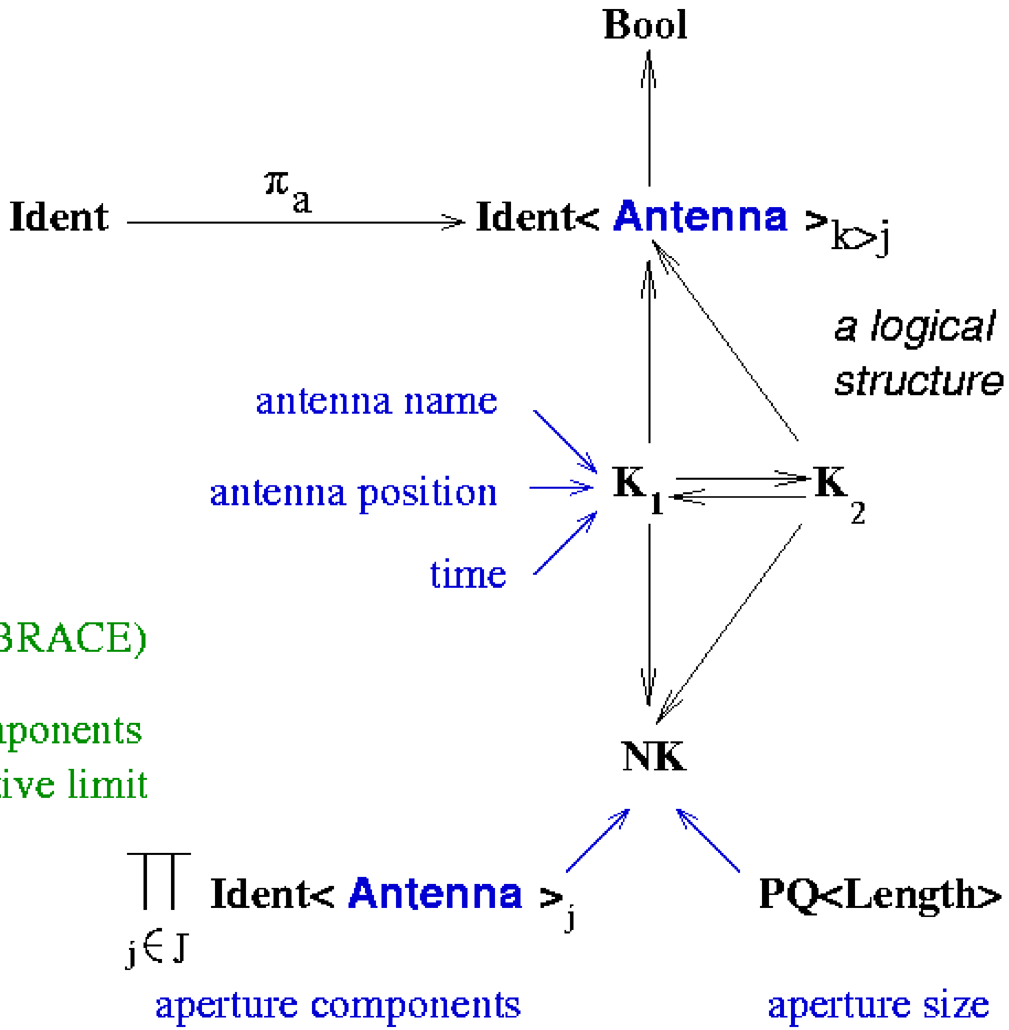
- Concepts which have contexts are topos:
  (→ keys are ordered sequences of foreign keys)
  (→ model not relational)

- The topology with 3 axes: aperture, frequency range and time range.

Relation Table< Antenna >



Relation Table< Antenna >

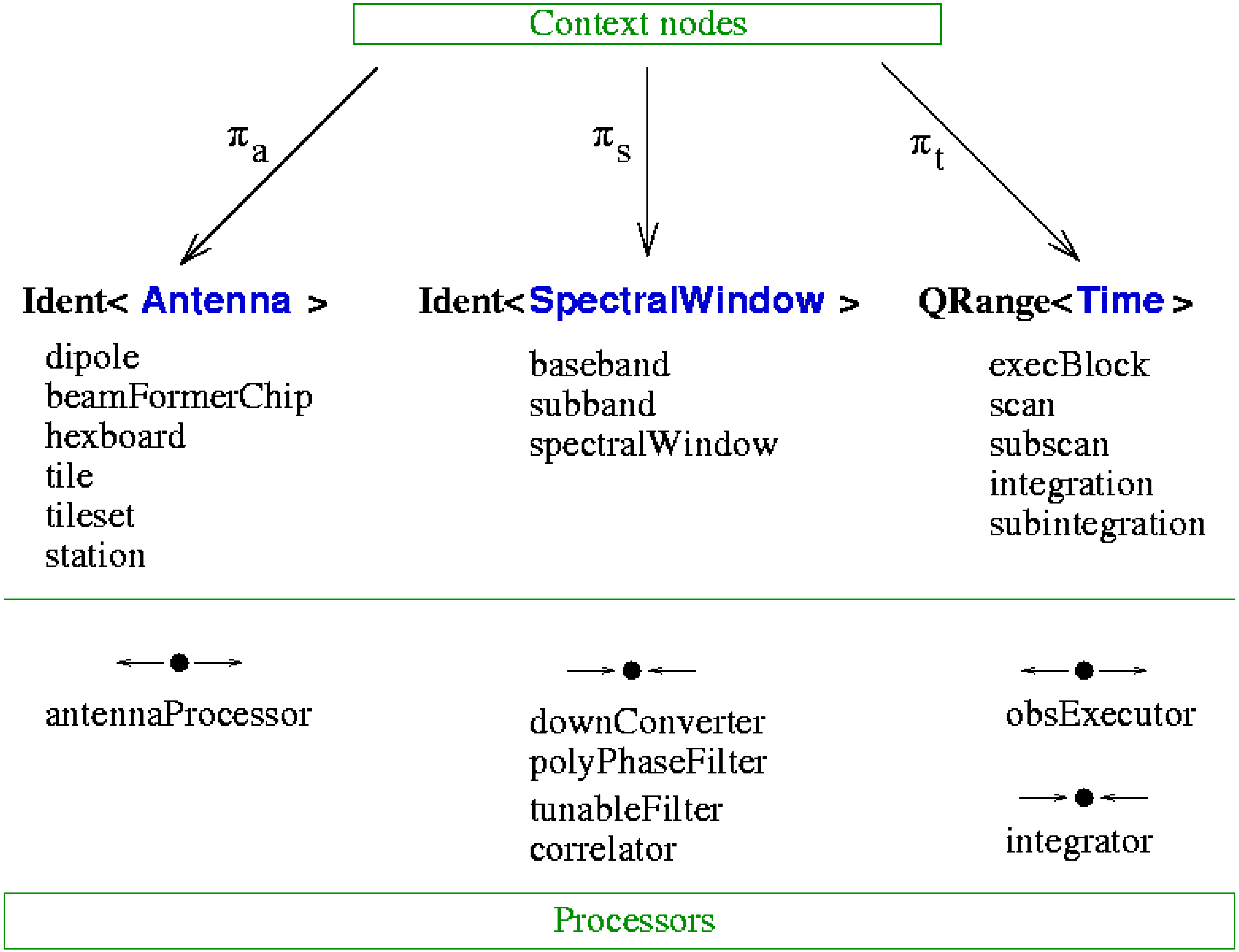


Use–case of APA (e.g. EMBRACE)

Antenna ⟶ Antenna components

A recursive object, a projective limit

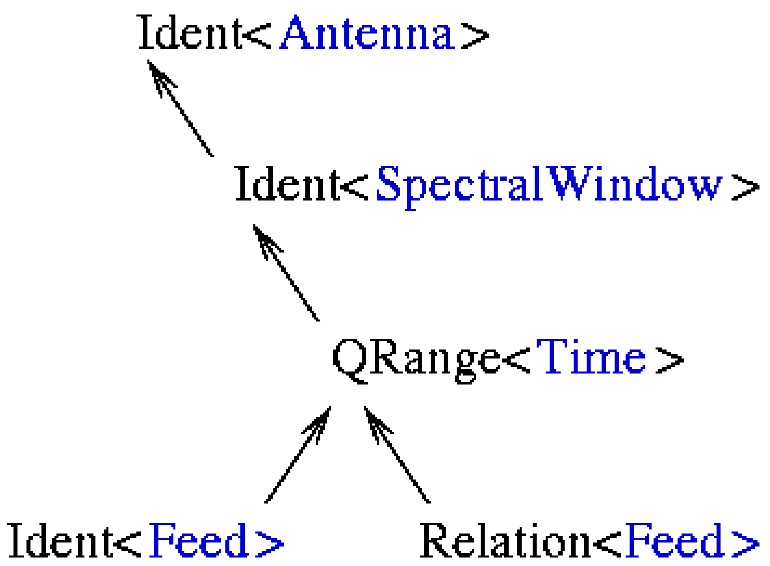
Topological space axis basis



Application: Table< Feed >

Proposition: A table is a monad which has for its algebraic structure a vector space, a directed set

Ident<Antenna> ⟶ Ident<SpectralWindow> ⟶ QRange<Time>



Conclusions

1. The theory of the measurement set has been mostly developed

2. The standard relational model is only a sub-category

3. Tables are sets containing a subset of their powersets, allow recursive definitions

4. Tables are monoids for $\uplus$

5. The Datset is a monoid

1. The formalism allows to support complex instruments such as aperture phased arrays

2. Generic programming in C++ allows to express this mathematical formalism: propotype SDMv2