What is “mapping” (supposed to do)?

Why again did we start this?

What can we agree on?

What really are the problems?

From the VO-DML spec (a standard!)

VO-DML is designed to satisfy the following requirements. It should

1. *Support the specification of serialization strategies for serializing instances of data models into different file formats;*
2. *Be rich enough to represent existing IVOA data models;*
3. *Support model reuse;*
4. *Be implementation-neutral, but...*
5. *Be flexible enough to be mapped to important physical representations, in particular XML schema, relational model (TAP), object-oriented languages (Java, Python...), and at the same time...*
6. *Be as minimal as possible, avoiding redundancy, adding restrictions where possible, with the aim of simplifying the work of modelers by offering few and "obvious" choices;*
7. *Be based on accepted standards for data modeling, but ...*
8. *Not rely on external modeling tools, but be sufficiently compatible with them so that such tools MAY be used when representing models;*
9. *Support runtime model interpretation;*

Some mapping use cases

http://wiki.ivoa.net/twiki/bin/view/IVOA/UtypesTigerTeam

Data Model (de)serialization

UC #1 Serialize DM instances to file: given an instance of a Data Model and the DM machine readable description, a writer can serialize the instance into a number of supported tabular formats. The writer could be a DAL service.

UC #2 Deserialize DM instance from file: given a serialized instance of a Data Model in a supported tabular format and the DM machine readable description, a reader can deserialize the instance into memory, building an object consistent with the DM itself.

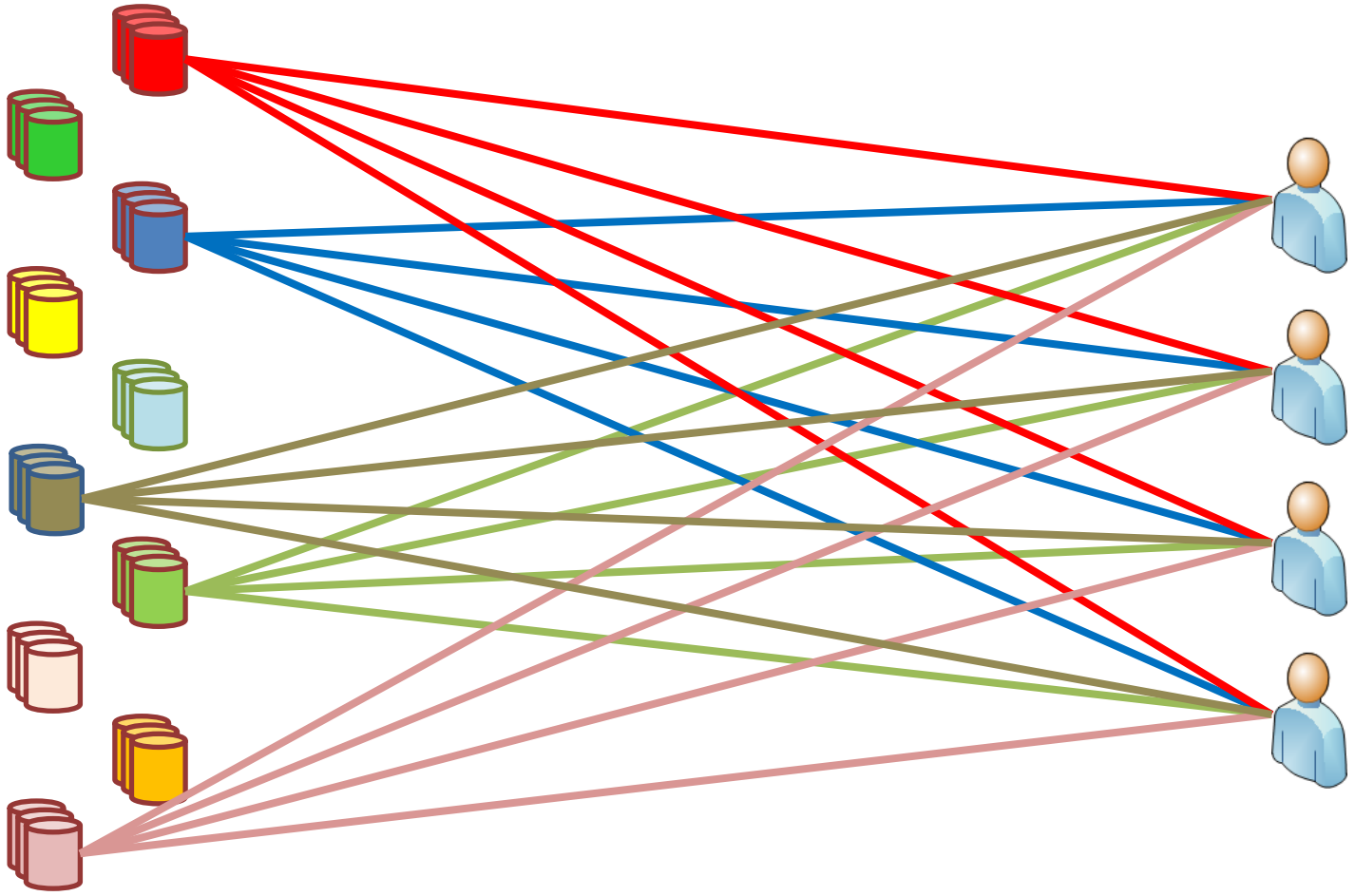
UC #3 Trivial round-tripping: given a serialized instance of a Data Model in a supported tabular format, an I/O library (possibly model-unaware) can convert the instance into a different, supported format without breaking its VO compliance.

UC #4 Represent an arbitrary number of instances of the same class in a DM instance (for example, N instances of the PhotometryFilter class in a PhotometryCatalog instance of the Spectral DM).

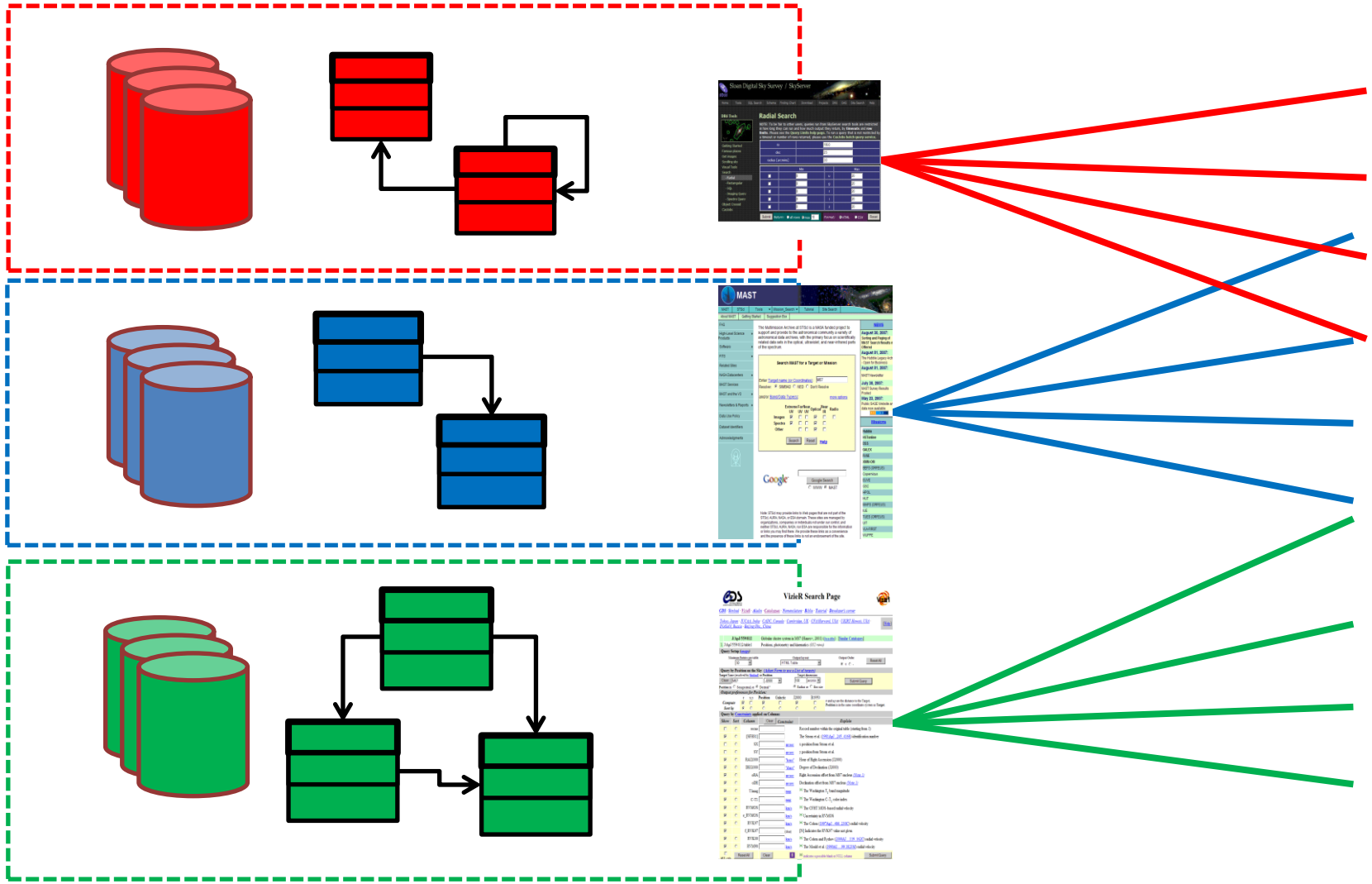
Why are we interested in data models?

Information Integration

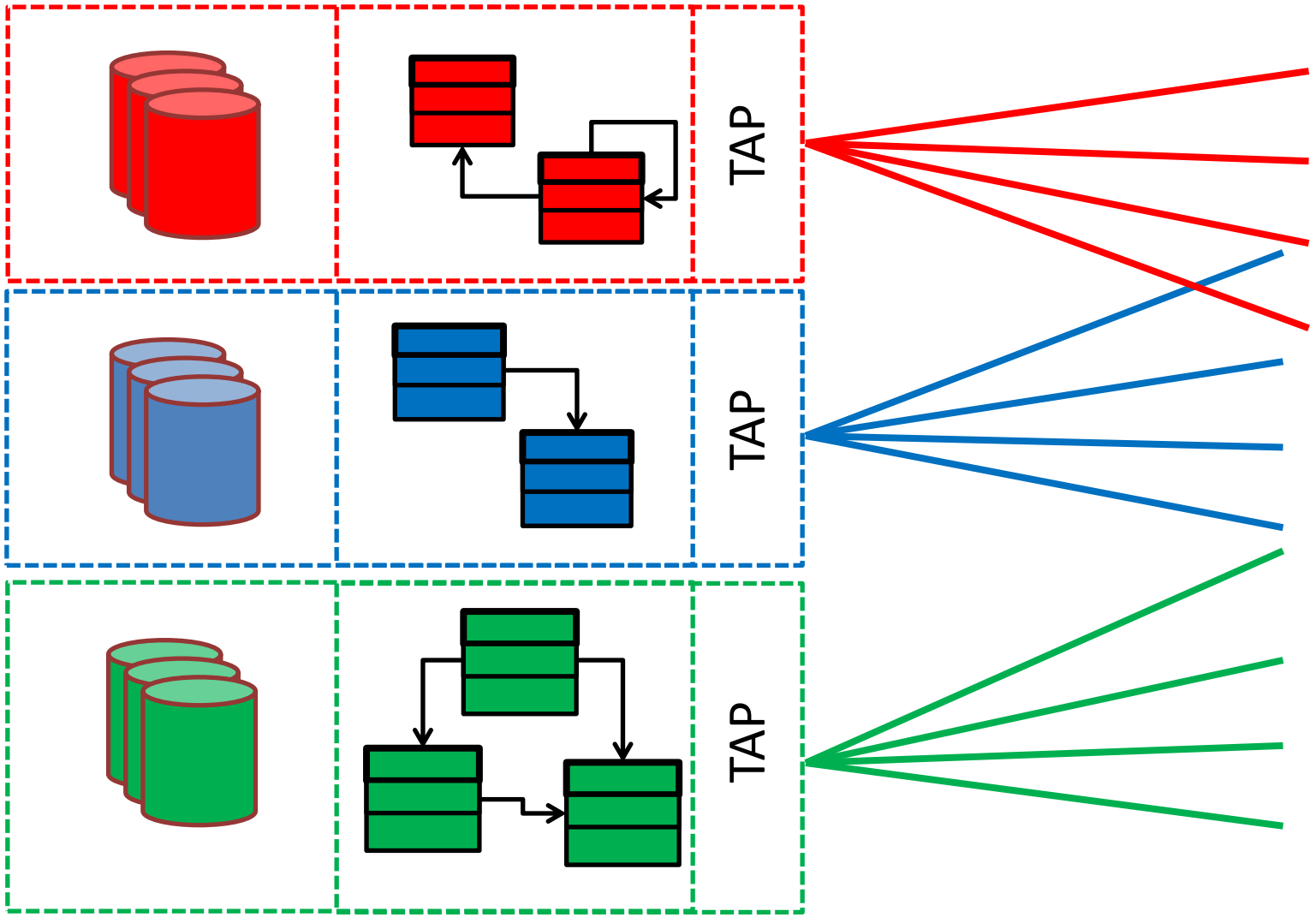
heterogeneity + scaling problem



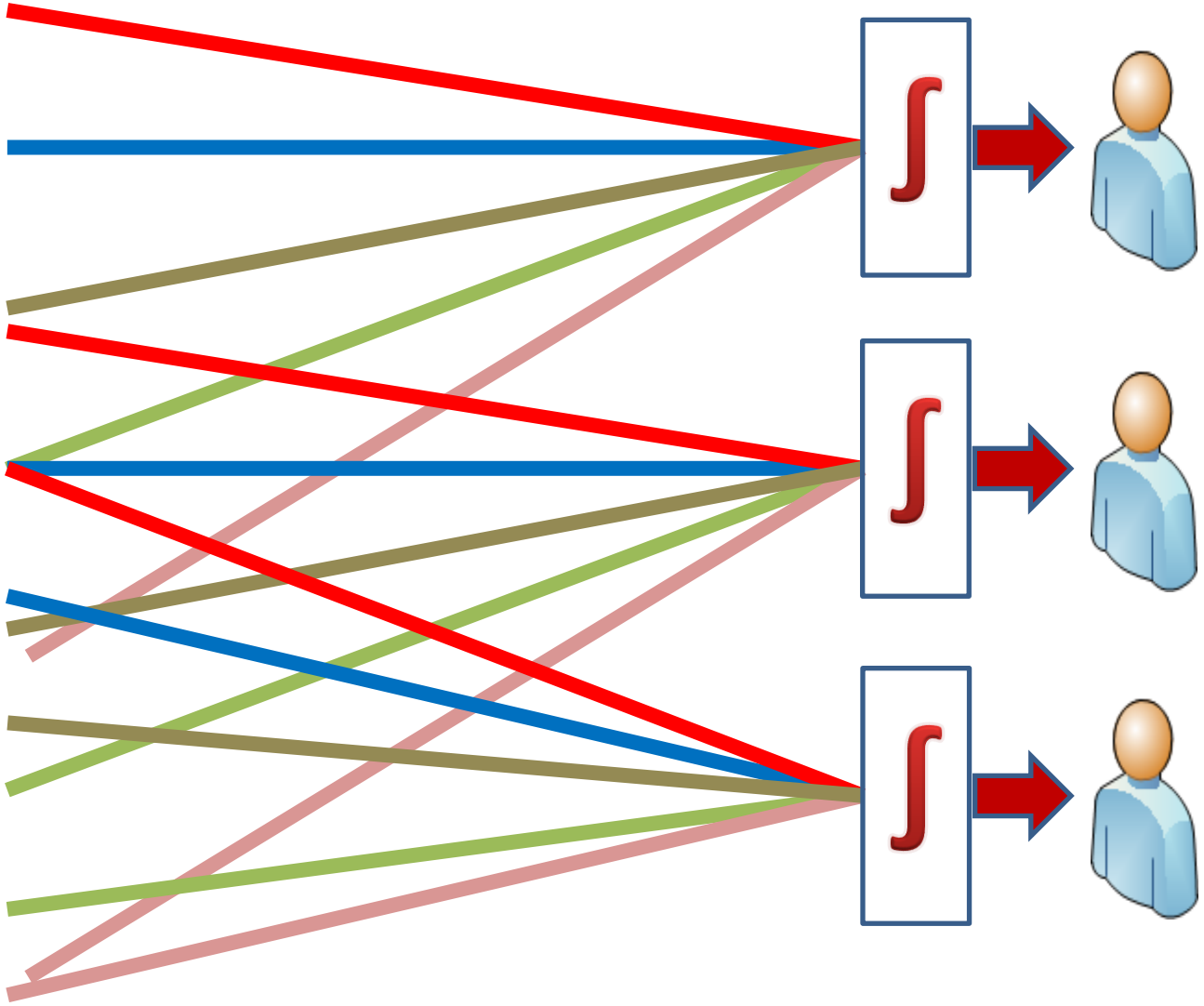
semantic: heterogeneous schemas
syntactic: custom access services



IVOA homogenizes syntax: e.g. TAP

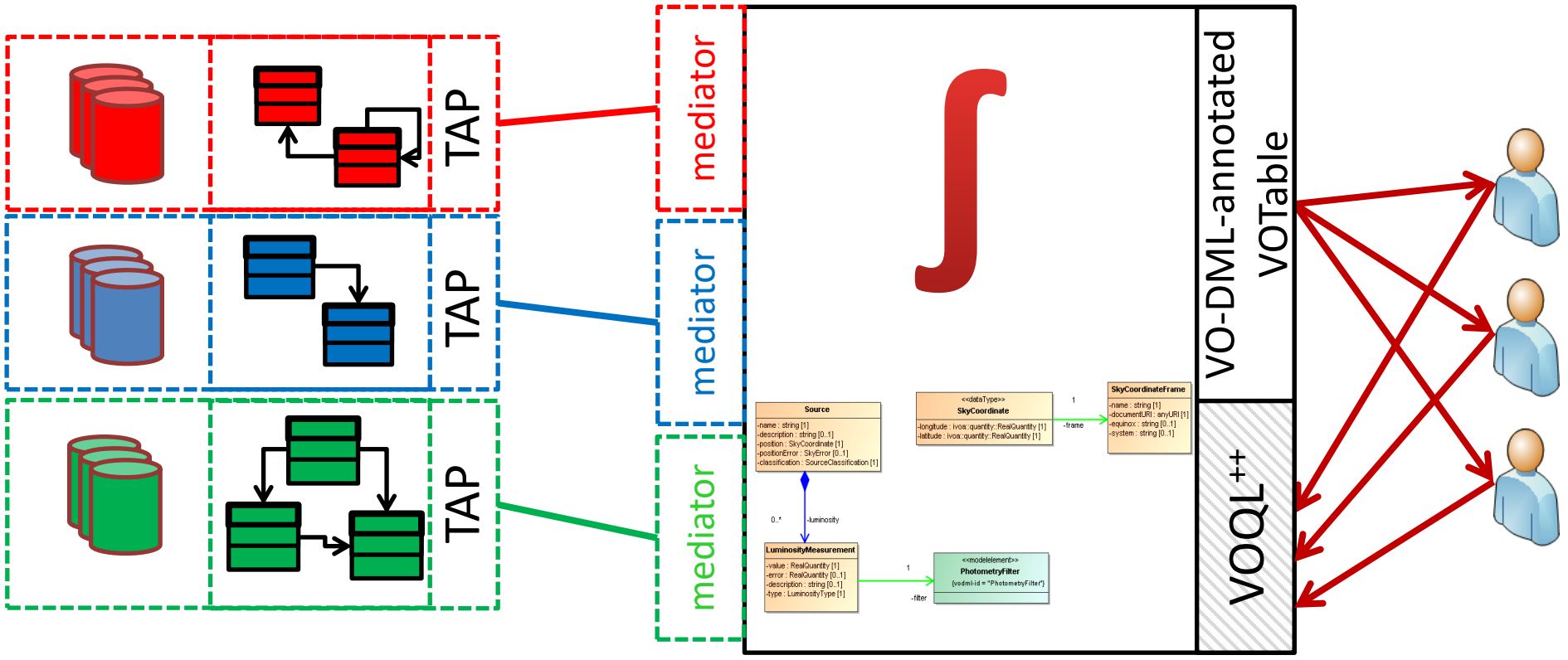


semantic heterogeneity requires individual data integration



Sketch of Integration Solution:
common (global) schema +
TAP + *mediation* + VOQL^{++}

(see lots of CS literature)



Global Schema(s)

==

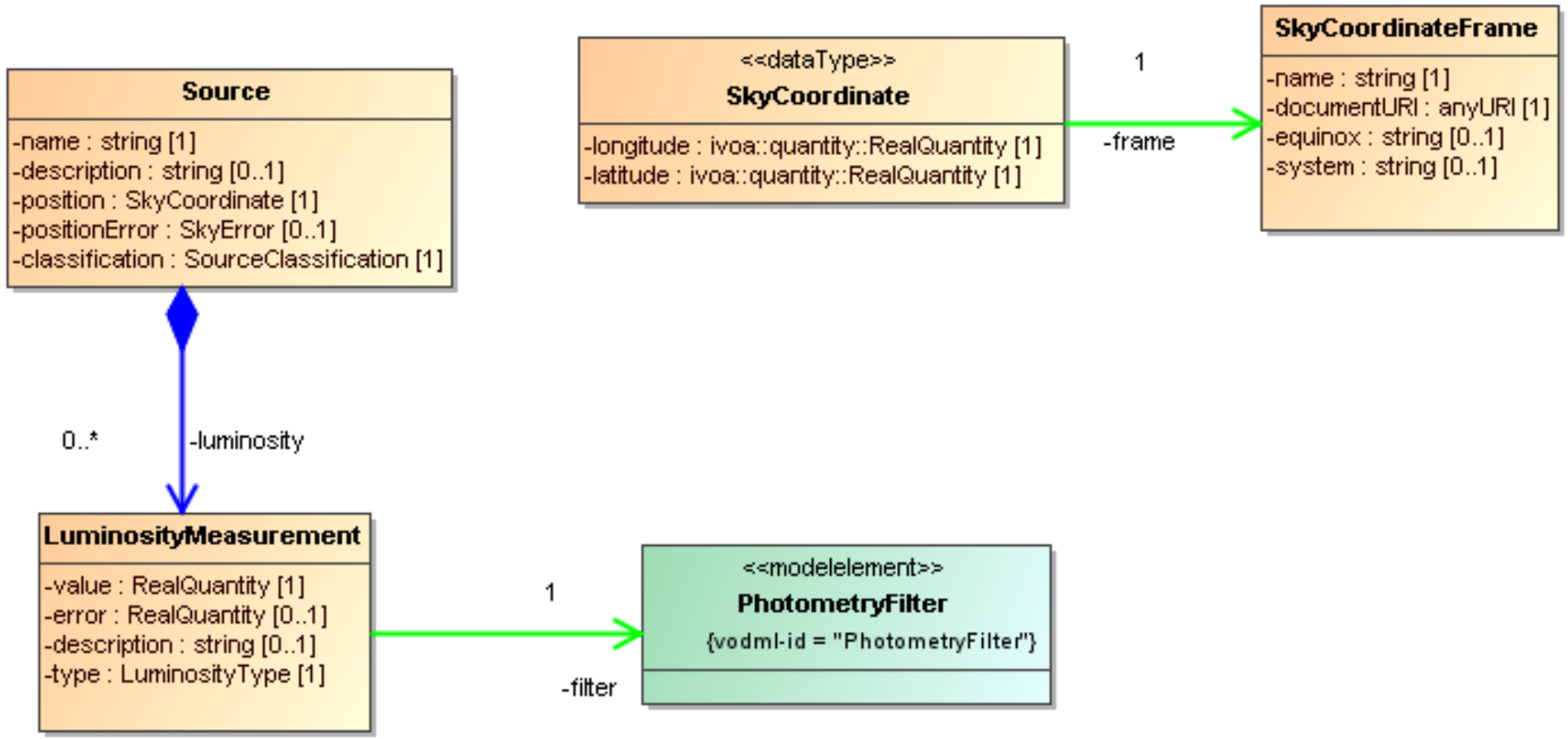
common data model(s)

- The *unified view* of the data sources
- Defined using VO-DML
  - supports model dependency/reuse
  - simplified, XML serialization language: machine readability!
  - Faithful representations possible
    - XSD
    - Java
    - YAML?

(VO-DML) mapping

- Expresses how instances of a data model (expressed as VO-DML) are represented in a tabular representation
  - VOTable
  - TAP schema

Example: Simple source data model



objid	ra	dec	u	g	r	i	z	run	rerun	camcol	field	specobjid	class	redshift	plate	mjd	fiberid
1237680191504712292	319.42017295	-2.91605515	19.453272	17.512213	16.43					5	45	4933710530549838848	GALAXY	0.091911	4382	55742	62
1237680191504842797	319.6726666	-2.89320328	18.006258	18.408295	17.02					5	47	4933700634945188864	STAR	-9.116632E-5	4382	55742	26
1237660241388240997	51.95792979	0.44178806	17.90674	16.767498	16.24					5	146	2329593891403098112	STAR	-2.435169E-4	2069	53376	389
1237660241388371981	52.15864799	0.5100779	18.619341	17.314531	16.76					5	148	2329602687496120320	STAR	-1.707261E-5	2069	53376	421
1237660241925505040	52.72667419	0.88746662	17.930399	16.900446	16.418163	16.214106	16.1061	3438	301	6	152	2329600488472864768	STAR	-1.622409E-4	2069	53376	413
1237660241925505156	52.85661769	0.97756273	18.178764	16.997499	16.512629	16.314194	16.207306	3438	301	6	152	2329599663839143936	STAR	-1.234436E-4	2069	53376	410
1237662303111507089	202.55299093	39.86892911	17.820675	16.164869	15.296254	14.812856	14.419583	3919	301	1	16	5299625250001449984	GALAXY	0.048569	4707	55653	52
1237663238729787997	52.05059022	0.14966321	19.351822	18.277271	18.06134	17.999191	17.999123	4136	301	4	165	2329595265792632832	STAR	-5.184785E-4	2069	53376	394
1237651271358408122	158.78373508	63.9613952	19.283352	17.41073	16.419657	16.042131	15.731997	1350	301	1	295	550602195343534080	GALAXY	0.11802	489	51930	135
1237651271358801125	158.82992158	63.94061555	19.297565	17.409573	16.431635	16.044048	15.707916	1350	301	1	295	550601370709813248	GALAXY	0.117888	489	51930	132

SDSS

Source

-name : string [1]
-description : string [0..1]
-position : SkyCoordinate [1]
-positionError : SkyError [0..1]
-classification : SourceClassification [1]

<<dataType>>
SkyCoordinate

-longitude : ivoa::quantity::RealQuantity [1]
-latitude : ivoa::quantity::RealQuantity [1]

1
-frame

-name : string [1]
-documentURI : any
-equinox : string [0..
-system : string [0..

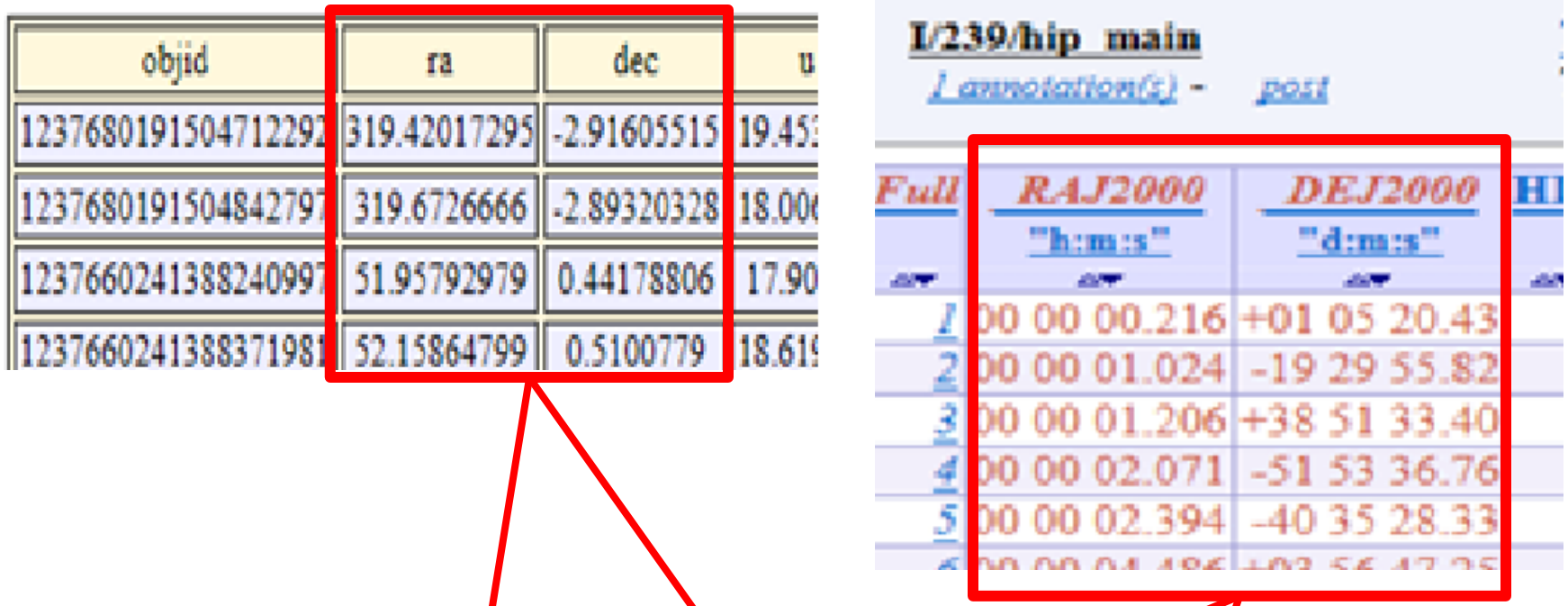
Identifying a table as containing Sources

I/239/hip_main
1 annotation(s) - *post*

The Hipparcos and Tycho Catalogues (ESA 1997)
The Hipparcos Main Catalogue *(118218 rows)*

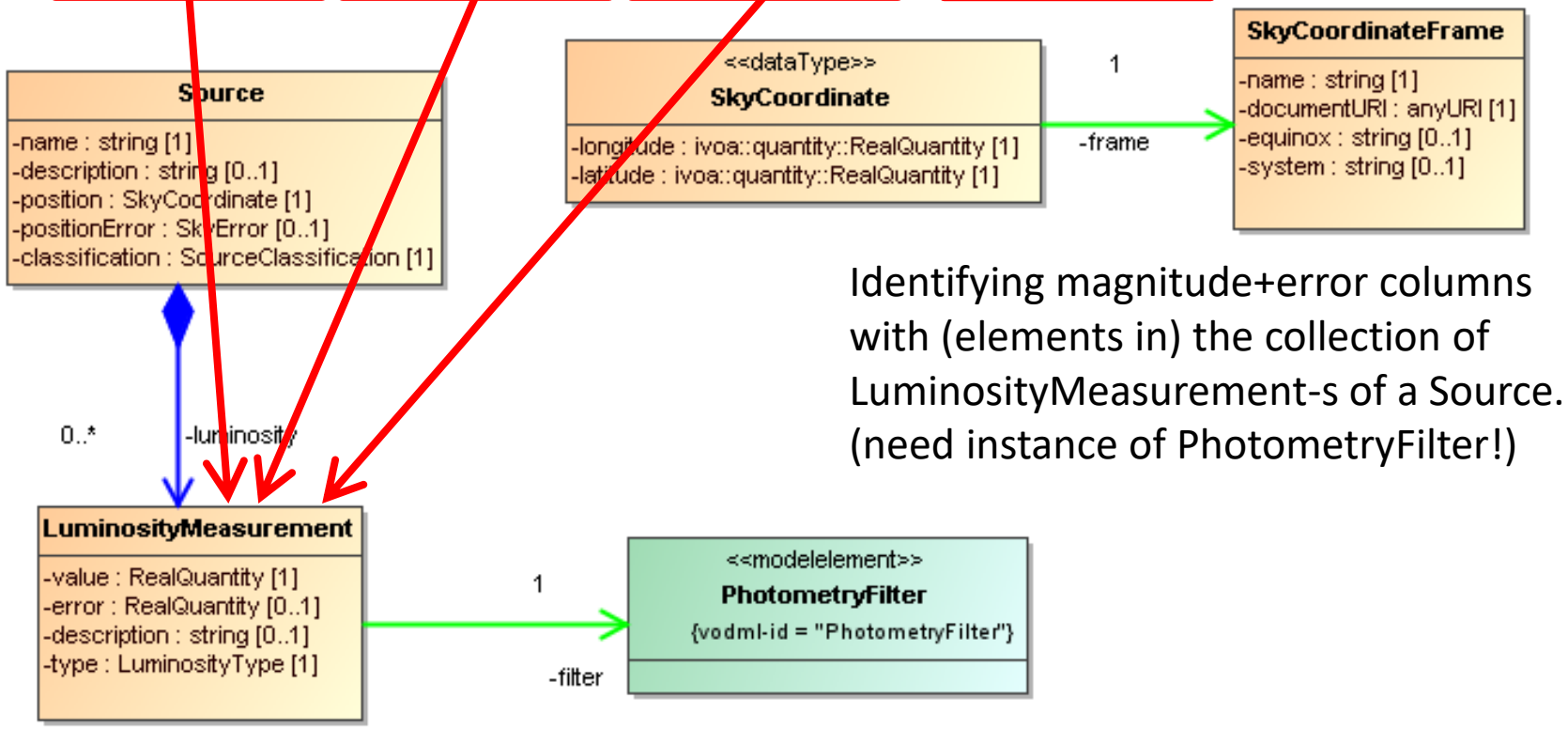
Full	RAJ2000 "h:m:s"	DEJ2000 "d:m:s"	HIP	RAhms	DEdms	Vmag mag	RA(ICRS) deg	DE(ICRS) deg	Plx mas	pmRA mas/yr	pmDE mas/yr	e_Plx mas	BTmag mag	e_ mag	VTmag mag	e_ mag	B-V mag	Hpmag mag	e_ mag
1	00 00 00.216	+01 05 20.43	1	00 00 00.22	+01 05 20.4	9.10	0.00091185	1.08901332	3.54	-5.20	-1.88	1.39	9.643	0.020	9.130	0.019	0.482	9.2043	0.0020
2	00 00 01.024	-19 29 55.82	2	00 00 00.91	-19 29 55.8	9.27	0.00379737	-19.49883745	21.90	181.21	-0.93	3.10	10.519	0.033	9.378	0.021	0.999	9.4017	0.0017
3	00 00 01.206	+38 51 33.40	3	00 00 01.20	+38 51 33.4	6.61	0.00500795	38.85928608	2.81	5.24	-2.91	0.63	6.576	0.004	6.621	0.005	-0.019	6.6081	0.0007
4	00 00 02.071	-51 53 36.76	4	00 00 02.01	-51 53 36.8	8.06	0.00838170	-51.89354612	7.75	62.85	0.16	0.97	8.471	0.007	8.092	0.007	0.370	8.1498	0.0011
5	00 00 02.394	-40 35 28.33	5	00 00 02.39	-40 35 28.4	8.55	0.00996534	-40.59122440	2.87	2.53	9.07	1.11	9.693	0.014	8.656	0.010	0.902	8.7077	0.0018
6	00 00 04.486	+03 56 47.25	6	00													1.336	12.4488	0.0085
7	00 00 05.283	+20 02 10.01	7	00									.542	0.039	9.679	0.030	0.740	9.6795	0.0021
8	00 00 06.562	+25 53 11.26	8	00									.433	0.055	9.151	0.029	1.102	8.5522	0.1671
9	00 00 08.477	+36 35 09.45	9	00									.962	0.025	8.711	0.015	1.067	8.7534	0.0018
10	00 00 08.740	-50 52 01.11	10	00									.140	0.011	8.630	0.010	0.489	8.6994	0.0020
11	00 00 08.961	+46 56 23.99	11	00 00 08.95	+46 56 24.0	7.34	0.03729695	46.94000154	4.29	11.09	-2.02	0.84	7.446	0.005	7.364	0.005	0.081	7.3777	0.0010
12	00 00 09.816	-35 57 36.81	12	00 00 09.82	-35 57 36.8	8.43	0.04091756	-35.96022482	4.06	-5.99	-0.10	1.16	10.369	0.023	8.588	0.010	1.484	8.5598	0.0012
13	00 00 10.008	-22 35 40.94	13	00 00 10.00	-22 35 40.9	8.80	0.04167970	-22.59468060	3.49	8.45	-10.07	1.48	10.216	0.026	8.887	0.014	1.128	8.9707	0.0017

Hipparcos@VizieR



Identifying two columns as the Position of a Source, a SkyCoordinate

ec	u	g	r	i	z	ru
505515	19.453272	17.512213	16.457823	15.911284	15.466995	808
320328	18.006258	18.408295	17.027491	16.594183	15.47355	808
78806	17.90674	16.767498	16.243202	16.038309	15.933592	343
00779	18.619341	17.314531	16.763399	16.53455	16.394312	343
46662	17.930399	16.900446	16.418163	16.214106	16.1061	343



Identifying magnitude+error columns with (elements in) the collection of LuminosityMeasurement-s of a Source. (need instance of PhotometryFilter!)

Seems straightforward no?

Mapping syntax?

This is what we've been discussing for years now.

So what's the problem?

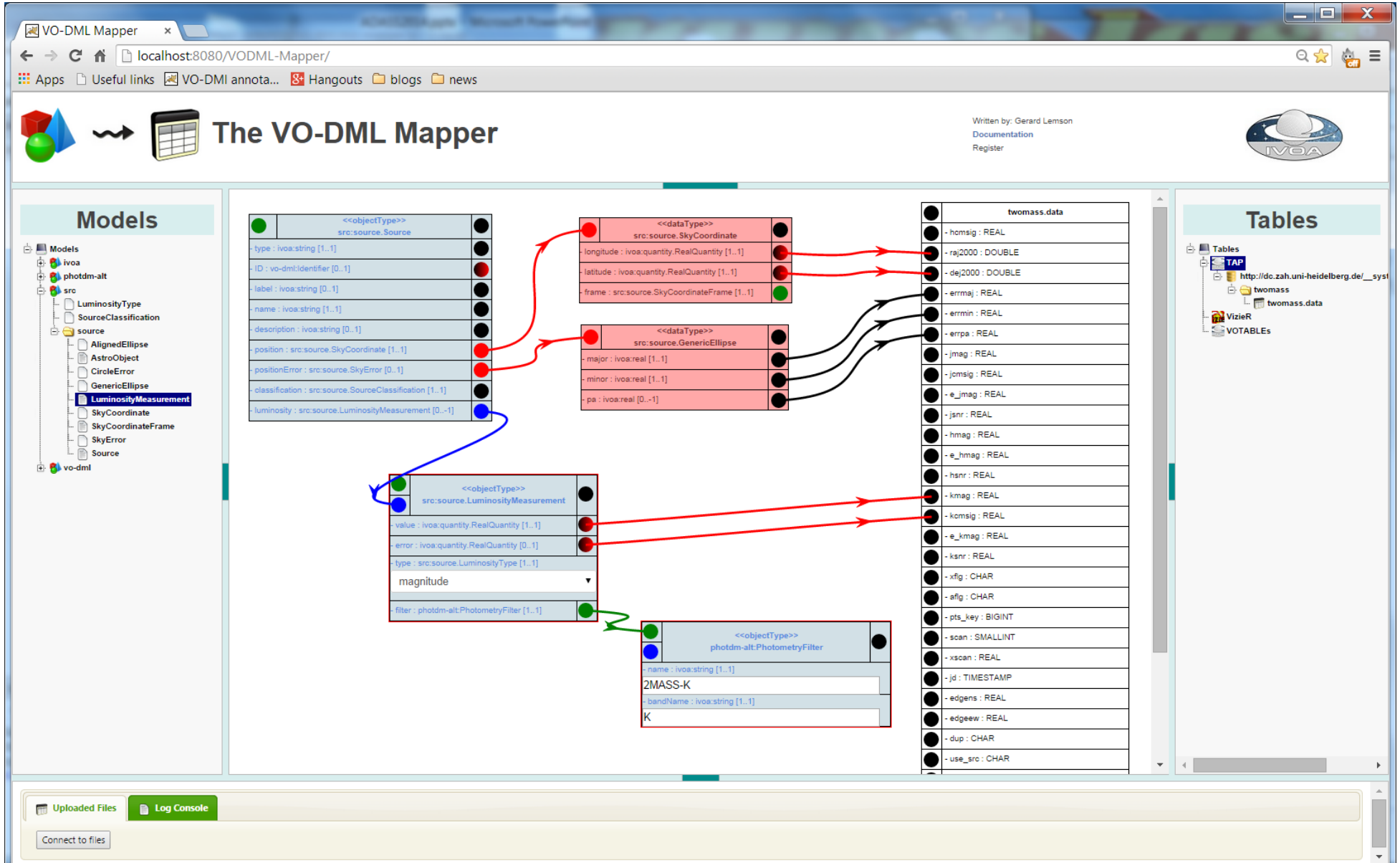
- Mapping *syntax*?
  - Maybe
- Or maybe mapping itself
  - Impedance mismatch between data models and data sets

component	serialization
Data model	VO-DML
Data set	VOTable, TAP
Mapping	(proposal)
Use cases	?

Graphical mapping tool

- allows mapping using drag-and-drop-and-connect from loaded models and TAP schemas/VOTables
- No syntax
- Shows impedance mismatch complexity

http://dsa012.pha.jhu.edu:8081/VODML-Mapper



Examples

Done.

History: utypes

- Introduced in VOTable 1.1
  - Together with GROUPing
- “pointer into data model”

- What were they?
- Tiger team 2012-> now

What is data model?
Why do we have them?

- Global schema in information integration

What to do with data models?

- Create instances, objects
  - In memory (code)
  - In database
  - In (XML) file
  - Conceptually, in our minds

How?

- *Map* data model to representation appropriate to usage

- Faithful:
  - Basically 1-1
  - E.g. ORM

Real world not so nice

- Legacy databases
- Query results

“what are utypes?”

translated to

“how do we identify DM instances in (VO)Tables?”

Ok?

So why has progress been slow?

- Even conceptually a HARD problem!

Impedance mismatch

Where is the problem?
3(4) components

component	serialization
Data model	VO-DML
Data set	VOTable, TAP
Mapping	?
Use cases	?

Conceptually complex

- Models don’t match
  - Hierarchy -> flat
  - Deep models -> “simple” data sets

Simple is in the eye of the beholder

- Table is simple if you want to pl0t col1 vs col2
- Not if you want to identify hierarchical model instances
  - *Translational semantics* in terms of pre-agreed-upon, “globa” schema

Where is the complexity?

- Data model? Often YES
- Data set? Generally simple
- Mapping?
  - Relation between data model and data set?
  - Syntax independent

Let’s have a look

- http://dsa012.pha.jhu.ed:8081/VODML-Mapper