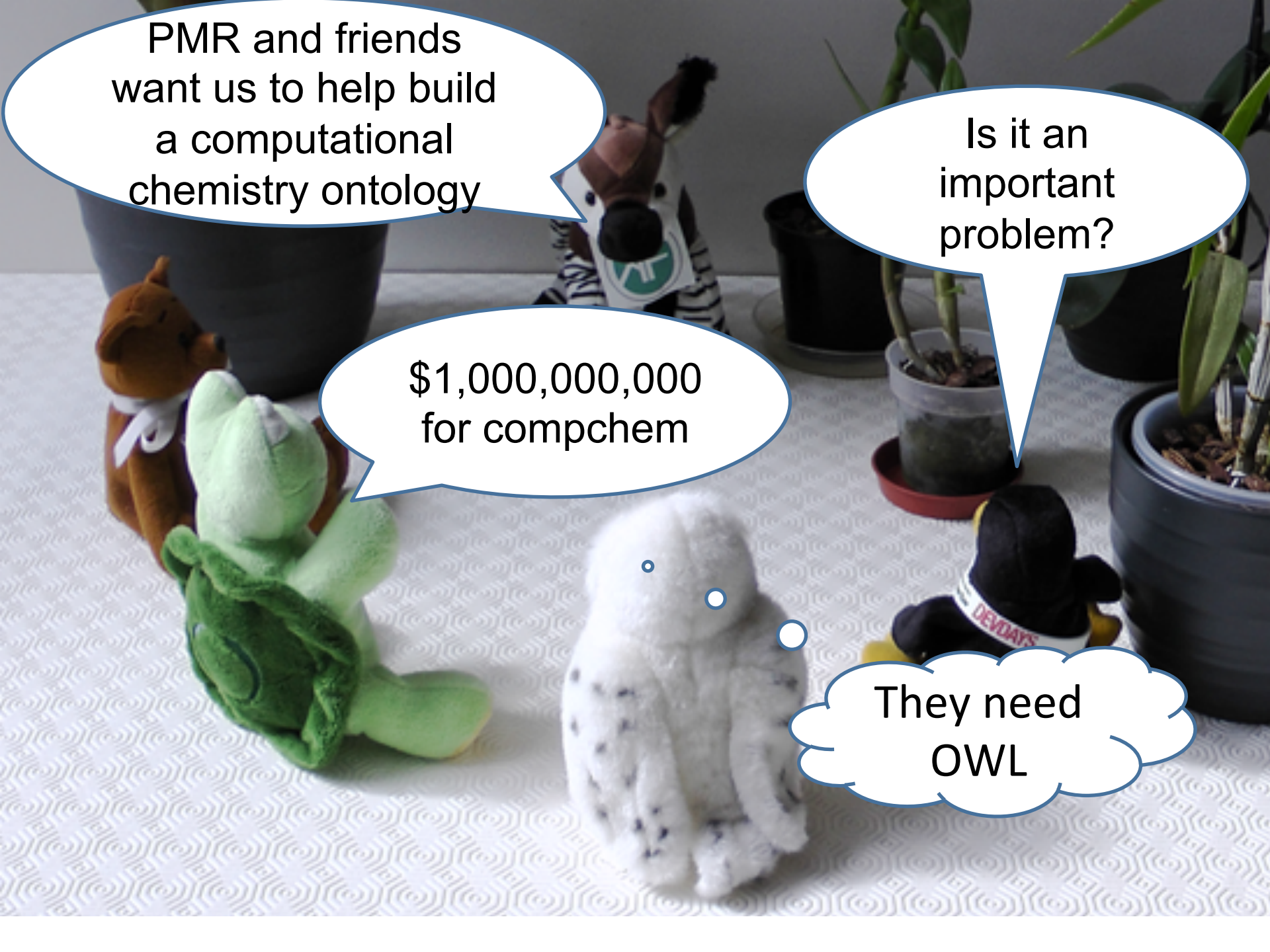


Ontologies in Physical Science

A photograph of a table covered with a white patterned cloth. On the table are several potted plants of various sizes and types, including a large dark grey pot with a green plant, a small black pot with a green plant, and a larger grey pot with a green plant. There are also several stuffed animals: a brown bear, a green frog, a white dog, a black penguin, and a zebra. The background is a plain grey wall.

Onto Workshop, ed.ac.uk 2013-04-11

An #animalgarden production and
Peter Murray-Rust, OKF & cam.ac.uk



PMR and friends
want us to help build
a computational
chemistry ontology

\$1,000,000,000
for compchem

Is it an
important
problem?


They need
OWL



Perhaps the
chemists could
use OWL-DL

Chemists don't
use ANY
ontologies

Top-down
schemas like
AniML haven't
(yet) taken off

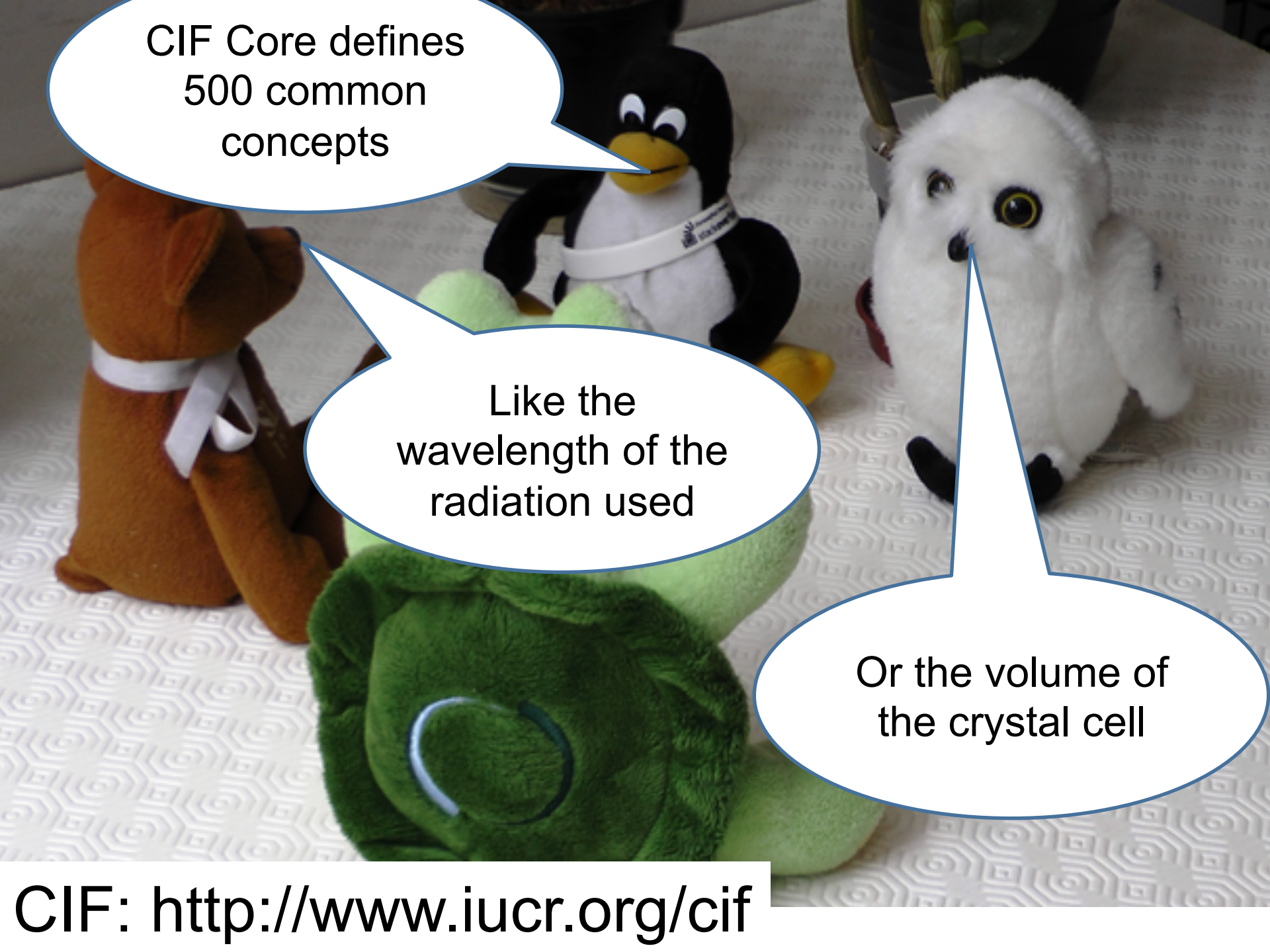


Are there any ontologies in physical science that work?

Crystallographers build CIF dictionaries

The IUCr, right? Tell us about CIF

IUCr: International Union of Crystallography



CIF Core defines
500 common
concepts

Like the
wavelength of the
radiation used

Or the volume of
the crystal cell

CIF: <http://www.iucr.org/cif>

An example ?

Core dictionary (coreCIF) version 2.4.3 _diffrn_ambient_temperature

Definition: The mean temperature in kelvins at which the intensities were measured.

Range: 0.0 -> infinity **Type:** numb

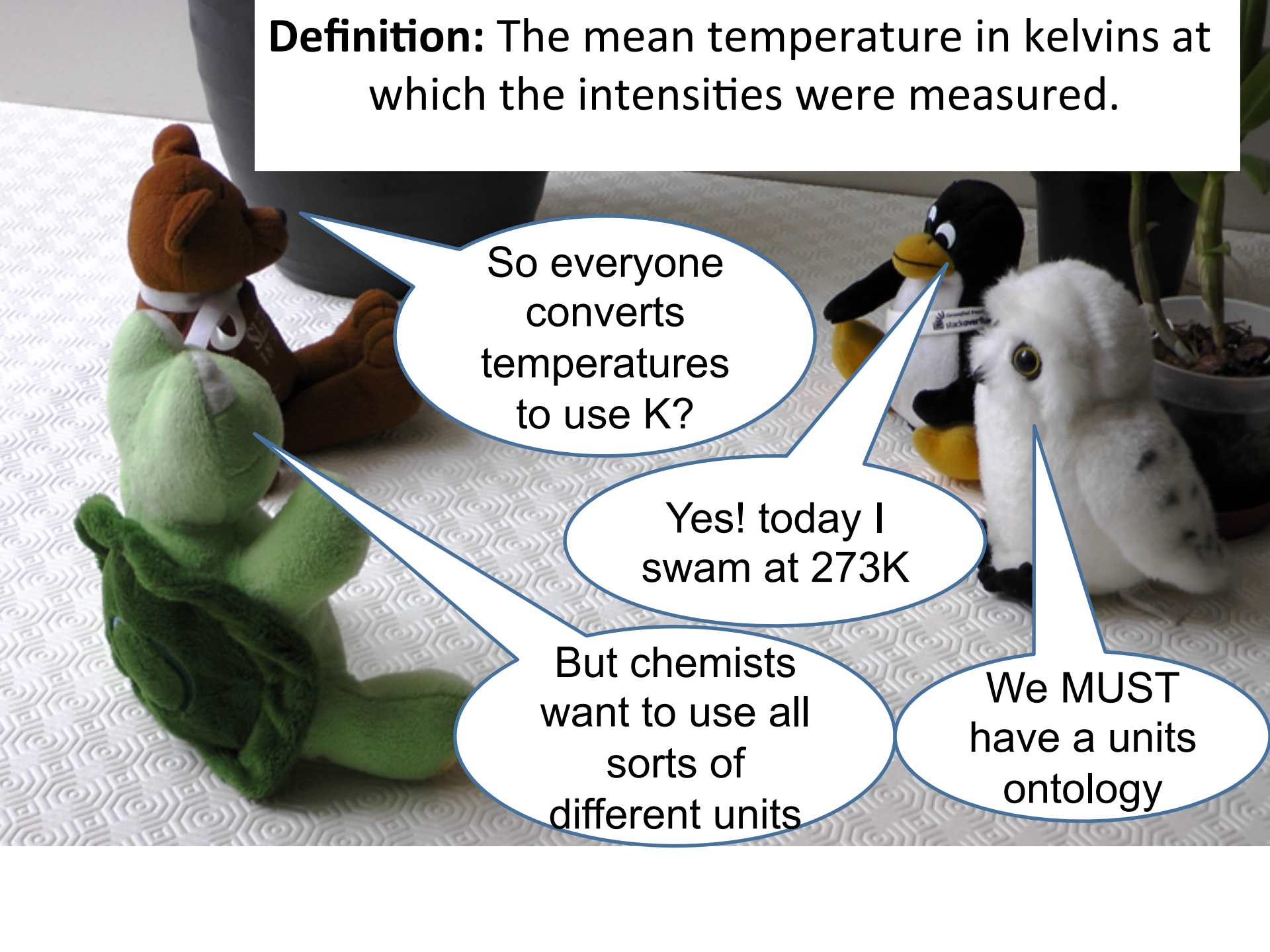
ID

For machines:
Constraint + type

For humans

http://www.iucr.org/data/iucr/cifdic_html/1/cif_core.dic/ldiffrn_ambient_temperature.html

Definition: The mean temperature in kelvins at which the intensities were measured.

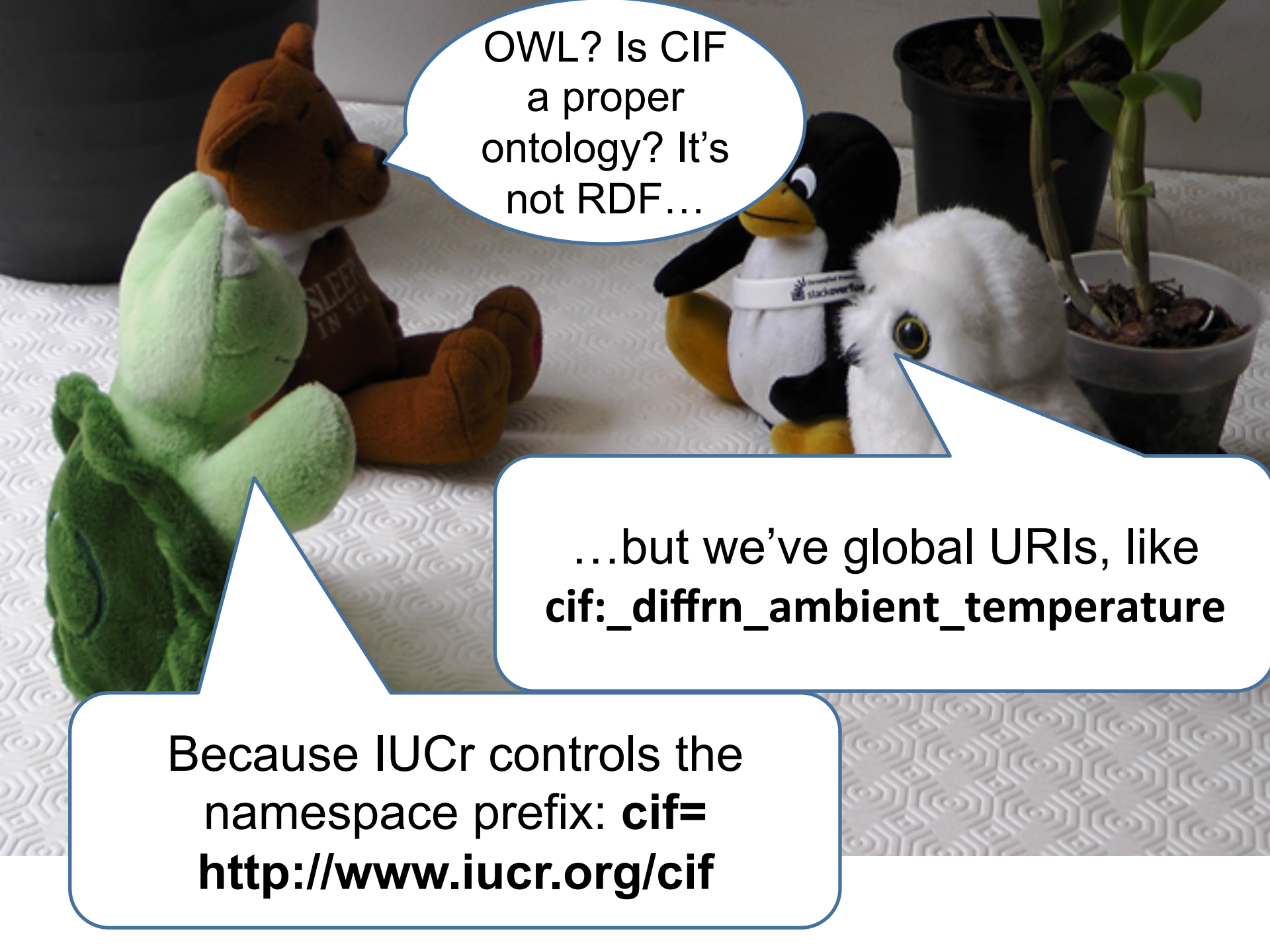


So everyone converts temperatures to use K?

Yes! today I swam at 273K

But chemists want to use all sorts of different units

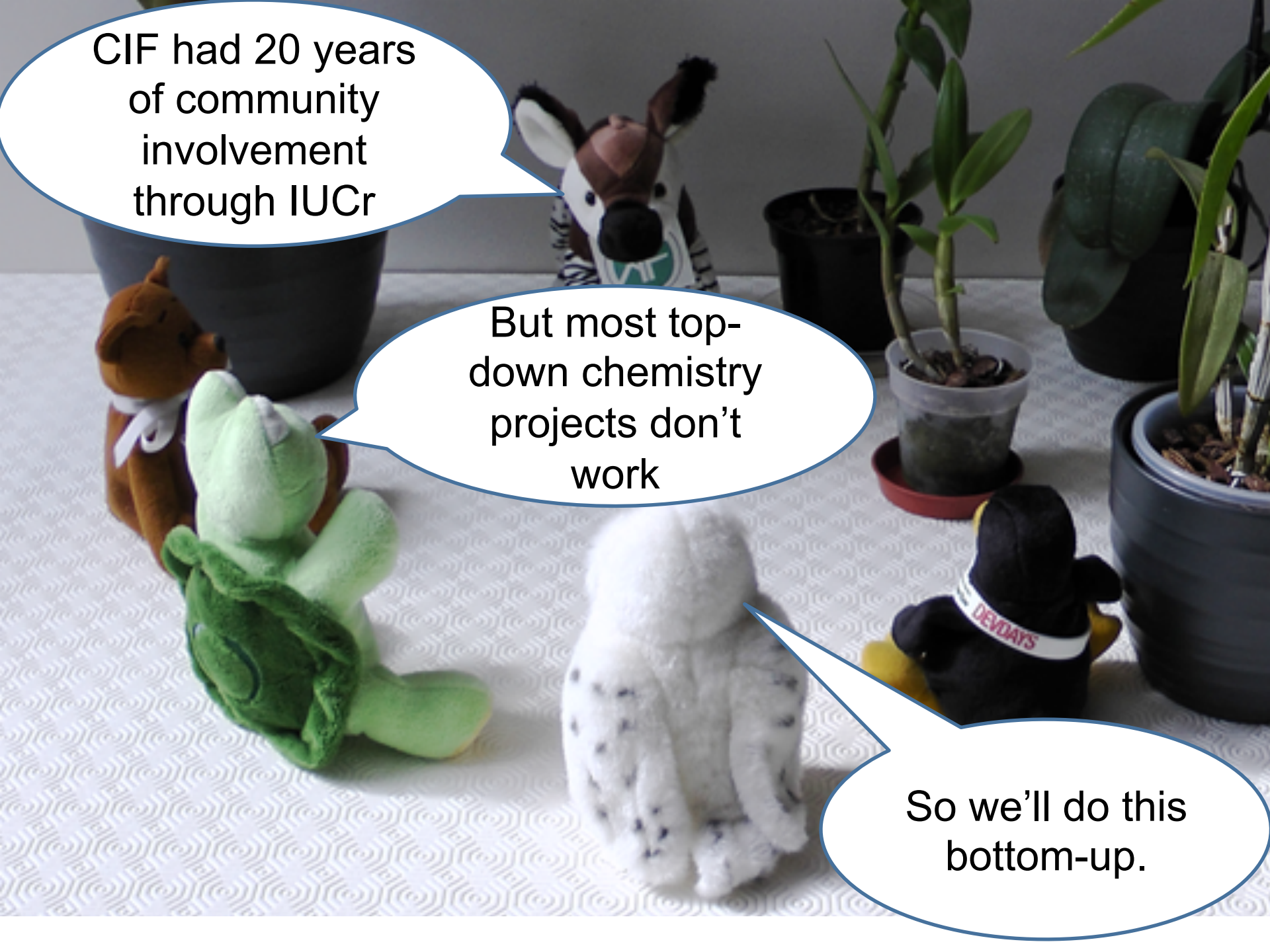
We **MUST** have a units ontology



OWL? Is CIF
a proper
ontology? It's
not RDF...

...but we've global URIs, like
cif:_diffn_ambient_temperature

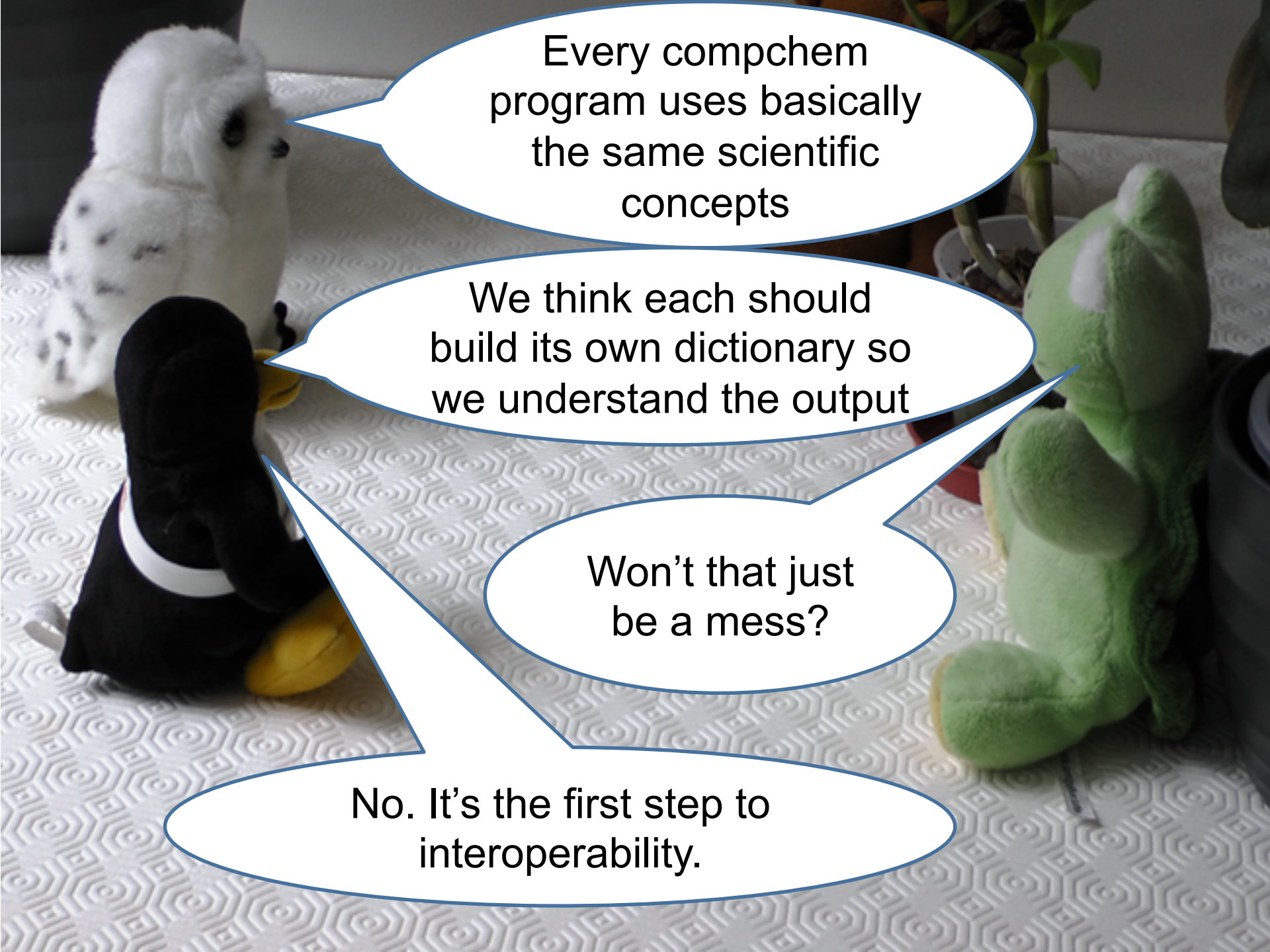
Because IUCr controls the
namespace prefix: **cif=**
<http://www.iucr.org/cif>

A collection of stuffed animals and potted plants on a white patterned surface. The stuffed animals include a brown dog, a green frog, a white dog with black spots, a black penguin with a red and white collar that says "DEV DAYS", and a brown and white zebra. There are several potted plants, including a green bamboo plant and a large green plant in a black pot. The background is a plain grey wall.

CIF had 20 years
of community
involvement
through IUCr

But most top-
down chemistry
projects don't
work

So we'll do this
bottom-up.

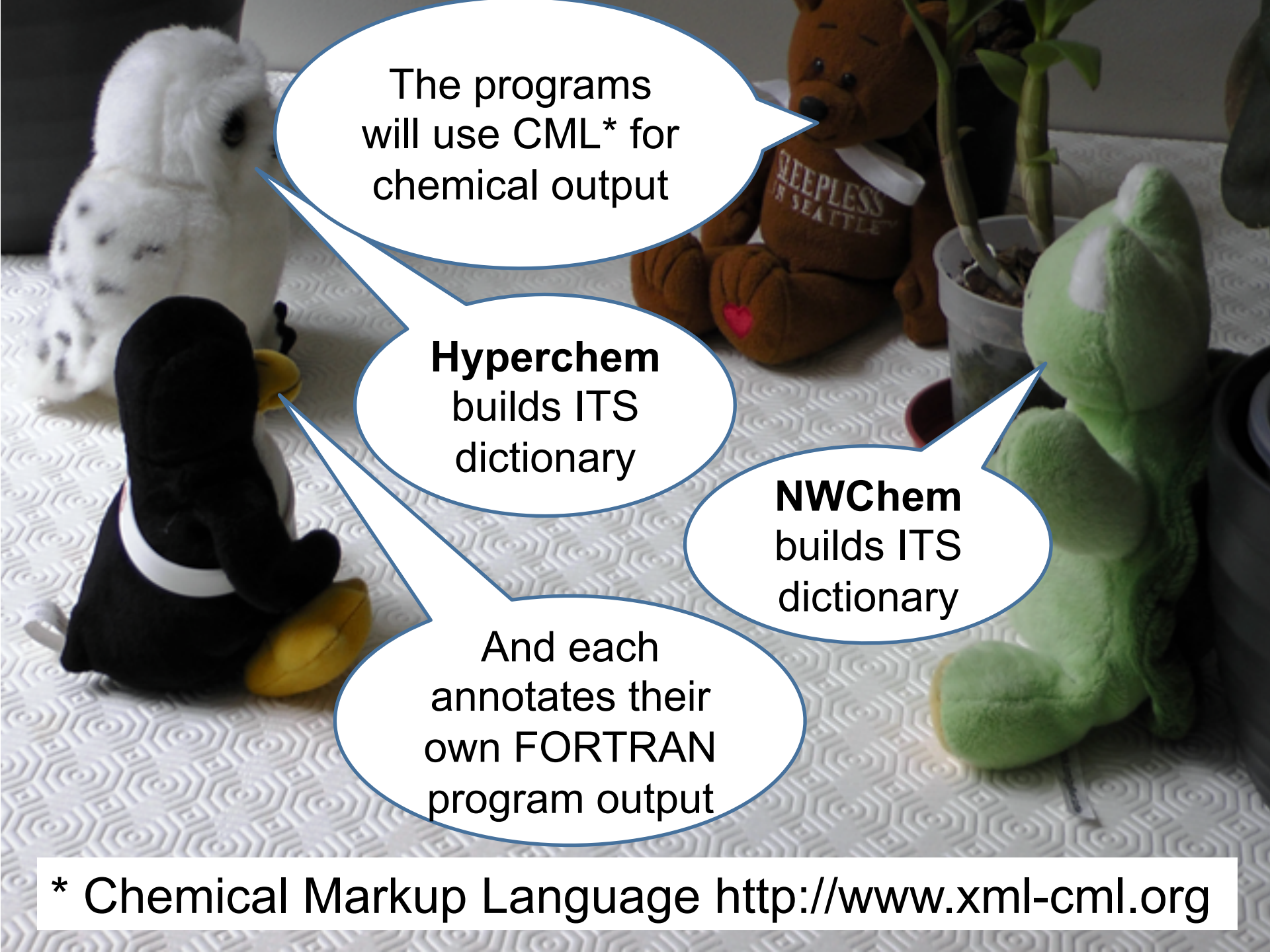


Every compchem
program uses basically
the same scientific
concepts

We think each should
build its own dictionary so
we understand the output

Won't that just
be a mess?

No. It's the first step to
interoperability.

The background of the slide features a collection of stuffed animals and plants on a white patterned surface. On the left, there is a white fluffy animal and a black and white penguin. In the center, a brown teddy bear is visible with the text 'SLEEPLESS IN SEATTLE' on its chest. To the right, there is a green frog-like stuffed animal and a small potted plant with green leaves.

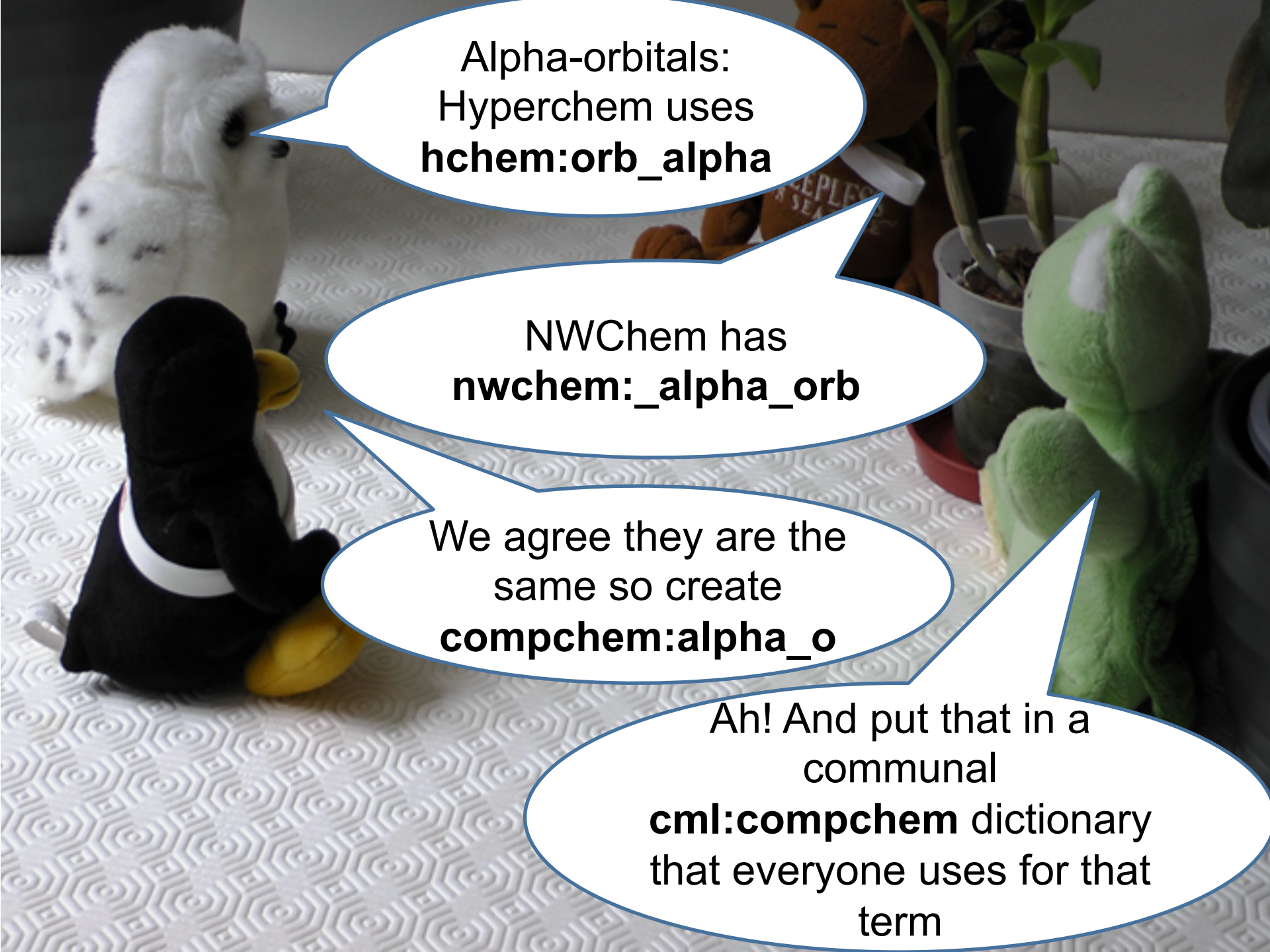
The programs
will use CML* for
chemical output

Hyperchem
builds ITS
dictionary

NWChem
builds ITS
dictionary

And each
annotates their
own FORTRAN
program output

* Chemical Markup Language <http://www.xml-cml.org>

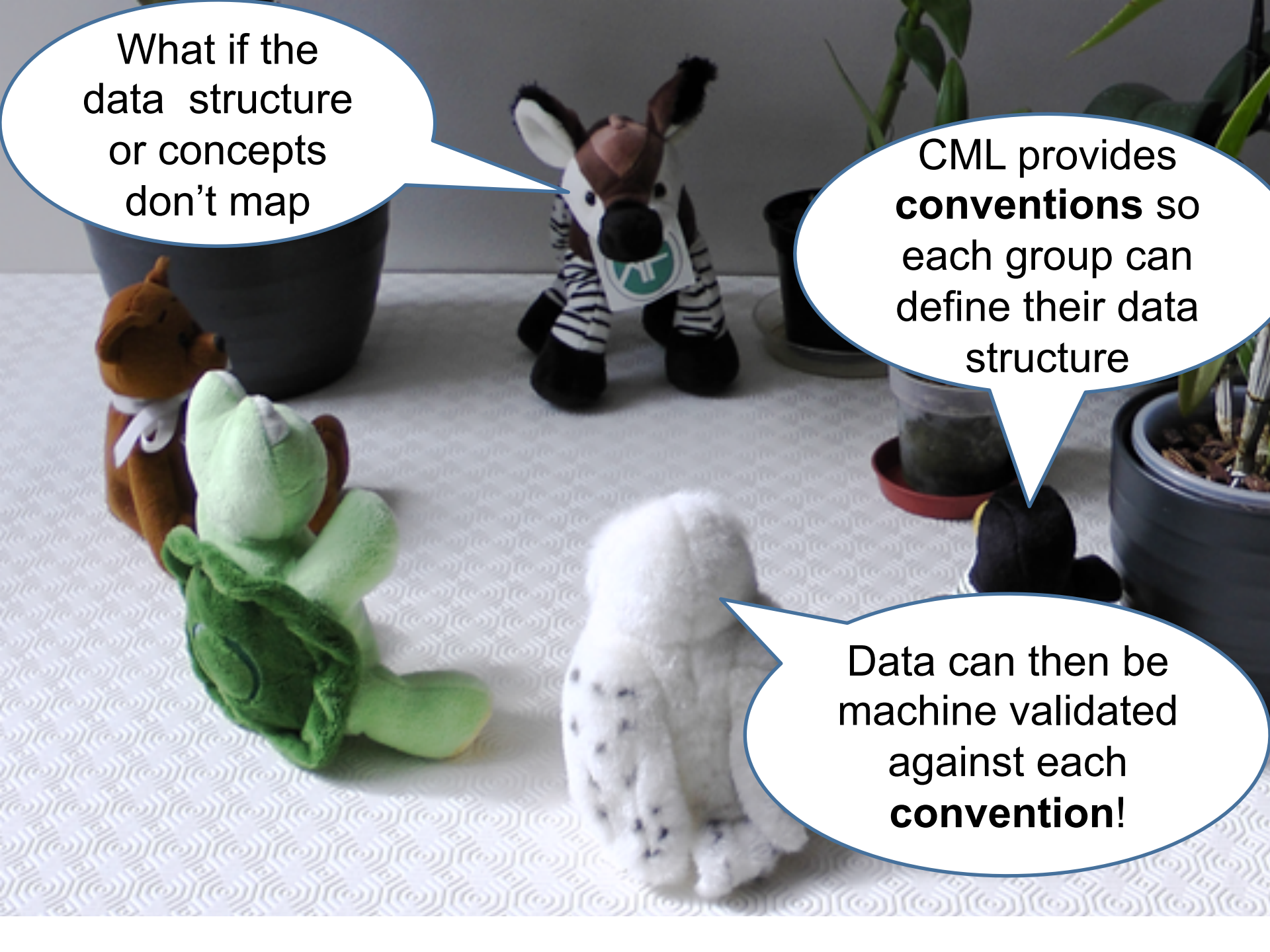
The background of the image shows a white patterned surface with several stuffed animals and plants. On the left, there is a white fluffy animal and a black and white penguin. On the right, there are green plants in pots. The text is overlaid on this scene in four speech bubbles.

Alpha-orbitals:
Hyperchem uses
hchem:orb_alpha

NWChem has
nwchem:_alpha_orb

We agree they are the
same so create
compchem:alpha_o


Ah! And put that in a
communal
cml:compchem dictionary
that everyone uses for that
term



What if the
data structure
or concepts
don't map

CML provides
conventions so
each group can
define their data
structure

Data can then be
machine validated
against each
convention!




But there are
over 20
program
codes.

We've
prototyped with
many before.
They'll be
encouraged

GULP,
DPOLY,
CASTEP,
SIESTA,
MOPAC ...

I think it's
going to work.
BUT TTT*

TTT: Things Take Time (Piet Hein)



Will it work? It depends on people

National labs
CSIRO/AU
and PNNL/US
are committed

I wish we had
some
publishers

And we have
companies like
Hyperchem
and Kitware

Benefits of semantic dictionaries:

- FORTRAN logfile can be made semantic
- High degree of interoperability in chemistry
- Semantic publication (HTML5, CML, MathML)
- Interoperates with mainstream Web
- Easily scalable to other phys sci.


Problems:

- Closed code/minds is short-term market advantage
- Non-trivial commitment (updates, code revision)
- Getting top-down approval (e.g. IUPAC)









What do we need?

We need
Dictionary
navigator+editor

We've got
FoX* for
FORTRAN
output

we've got Jumbo
Templates for
parsing logfiles



