Literate serialization of linguistic metamodels

Piotr Bański

Institut für Deutsche Sprache, Mannheim;

Institute of English Studies, University of Warsaw

Balisage: the Markup Conference, 4 August 2011

Literate serialization of linguistic metamodels

or: can one half-baked standard save another?

Piotr Bański

Institut für Deutsche Sprache, Mannheim;

Institute of English Studies, University of Warsaw

Balisage: the Markup Conference, 4 August 2011

Bird's eye view of the talk

This presentation looks at a part of the actual implementation attempt behind a long theoretical story that I have been working on for a few months.

TOC:

- Visual modelling as used in (some) ISO TC37 SC4 work on LR standardisation
- My project, involving TEI (TEI is cool)
- Visual modelling of XML applications
- TEI again (it's cool)
- An idea and a bit of musing

Visual modelling in some ISO work

ISO Committee TC37 SC4 (http://www.tc37sc4.org/)
doing great job on issues of language-resource management

My example: ISO LMF (Lexical Markup Framework, ISO 24613:2008)

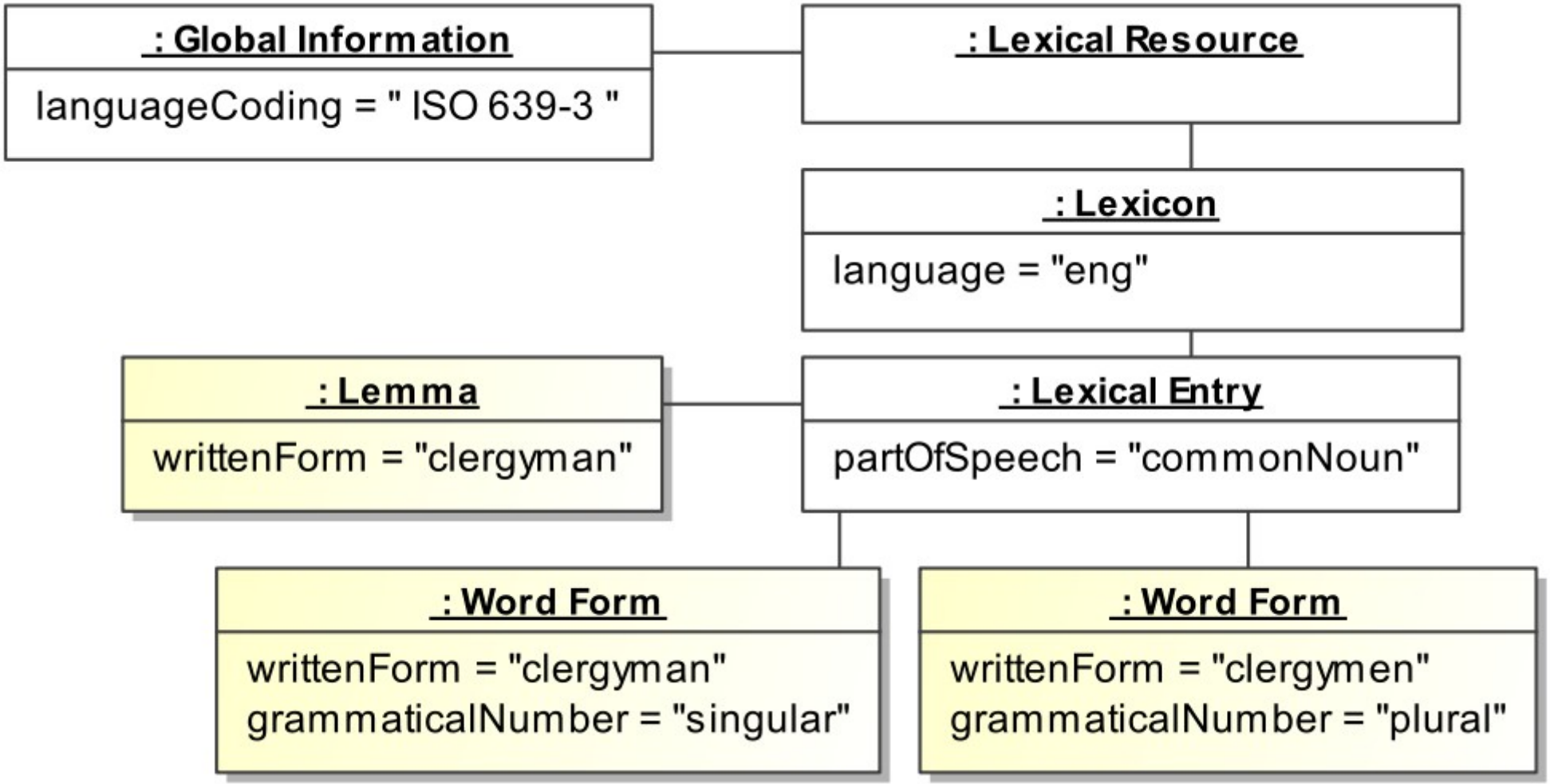
There's a lot of boxes there...

… and they are interlinked in various ways.

It's meant as a really cool system, whereby you are supposed to put various pieces together to obtain... larger pieces that do stuff!

Look:

Example LMF instance diagram (from Wikipedia)

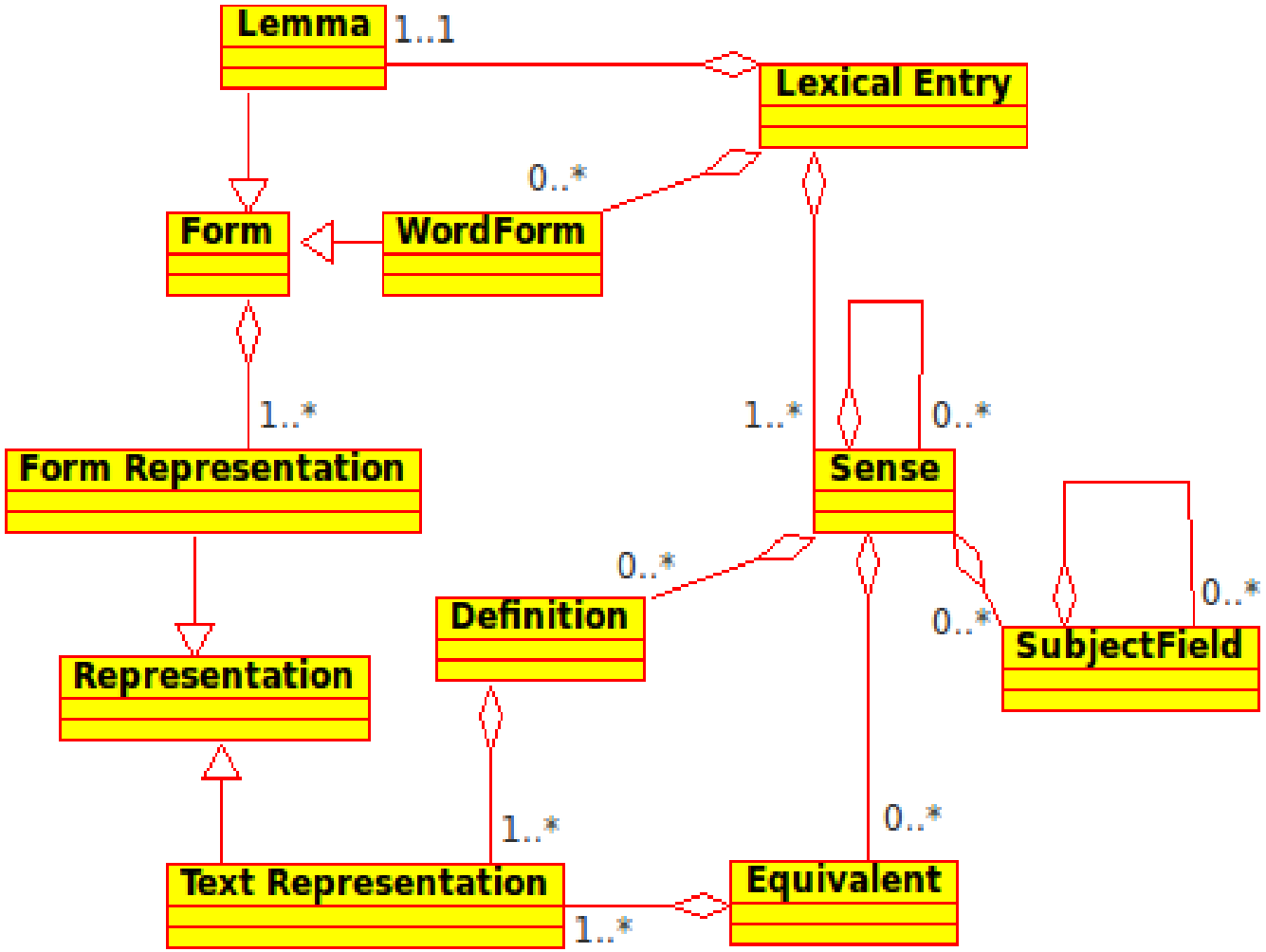


Example serialization (also from Wikipedia)

```
<LexicalResource dtdVersion="15">
    <GlobalInformation>
        <feat att="languageCoding" val="ISO 639-3"/>
    </GlobalInformation>
    <Lexicon>
        <feat att="language" val="eng"/>
        <LexicalEntry>
            <feat att="partOfSpeech" val="commonNoun"/>
            <Lemma>
                <feat att="writtenForm" val="clergyman"/>
            </Lemma>
            <WordForm>
                 <feat att="writtenForm" val="clergyman"/>
                 <feat att="grammaticalNumber" val="singular"/>
            </WordForm>
            <WordForm>
                <feat att="writtenForm" val="clergymen"/>
                <feat att="grammaticalNumber" val="plural"/>
            </WordForm>
        </LexicalEntry>
    </Lexicon>
</LexicalResource>
```

A fragment of LMF core model

If you need a simple lexicon, you choose a simple set of structural boxes; if you need to describe the morphological, syntactic or semantic properties of words or even of relationships among words, you go for a more complicated set of boxes...



WordForm
+ tense :
+ person :
+ number :
+ writtenForm :
+ phoneticForm :

tense, person and number coming from ISO-DCR)

Further info on ISO LMF

- http://www.lexicalmarkupframework.org/
- Work on the LMF took 5 years, and
- the result is “a synthesis of the state of the art in NLP lexicon field”.
- Having been accepted in 2008, it is now a 3-year-old **standard**.

Further info on ISO LMF

- http://www.lexicalmarkupframework.org/
- Work on the LMF took 5 years, and
- the result is “a synthesis of the state of the art in NLP lexicon field”.
- Having been accepted in 2008, it is now a 3-year-old **standard**.

This standard has possibly around 3 implementations, per the homepage.

Further info on ISO LMF

- http://www.lexicalmarkupframework.org/
- Work on the LMF took 5 years, and
- the result is “a synthesis of the state of the art in NLP lexicon field”.
- Having been accepted in 2008, it is now a 3-year-old **standard**.

This standard has possibly around 3 implementations, per the homepage.

Or indeed rather one:
Lexus; MPI Nijmegen, “the first test and reference implementation of LMF.” http://www.lat-mpi.eu/tools/lexus/lexus-description

Further info on ISO LMF

- http://www.lexicalmarkupframework.org/
- Work on the LMF took 5 years, and
- the result is “a synthesis of the state of the art in NLP lexicon field”.
- Having been accepted in 2008, it is now a 3-year-old **standard**.

This standard has possibly around 3 implementations, per the homepage.

Or indeed rather one:
Lexus; MPI Nijmegen, “the first test and reference implementation of LMF.” http://www.lat-mpi.eu/tools/lexus/lexus-description

Well, actually, a half: Lexus implements only two classes (Form and Sense), implies two others, and leaves the rest to be constructed by hand, for the time being (Menzo Windhouwer, p.c.); it's a good idea and actually a promising system, so I keep my fingers crossed.

Why are things so bad? (I)

There is a pressure for standardization of LRs, but there are also other standards on the market:

- OLIF (dying out, possibly dead);
- ISO 1951 – meant for print dictionaries; apparent failure, except for the publishing house that has financed it;
- SIL LIFT (specific **interchange** standard with serious tool applications and UML modelling behind it; very promising);
- TEI Dictionary module – hardly a standard, unless seen in the context of the entire TEI;
- others.

Why are things so bad? (II)

LMF gets lots of PR, I've read some 15 papers dealing directly with it.

This is good, but apparently not enough. Something's missing.

- Lack of tools,
- the serialization syntax does not appear to be worth much (a hand-crafted DTD), and Lexus may actually have its own serialization dialect;
- Lexus has a community of field linguists, but field linguists aren't after standards, XML and modelling. They are chasing vanishing languages, and they want to do that efficiently. SIL has a lot to offer in this respect.

But don't get me wrong

I respect the work put into ISO LMF, and I believe that, with a few fixes, it should be able to get off the ground.

Some shortcomings (which UML?, aggregation vs. composition, mapping issues, internal issues in the model, e.g. lack of constraints), but come on... everything can be improved.

Lack of **accepted** concrete syntax for the model (issue voiced strongly by Laurent Romary, I am following his suggestion in my ISO-TEI work).

There is also a fertile basis for development (FreeDict.org, which I happen to co-admin: some 72 dictionaries, almost all of them encoded in TEI P5).

Project description (aaabridged)

Supply ISO LMF with concrete syntax; **apply** the results immediately.

Concrete syntax – any ideas?

TEI Dictionaries (ch. 9): started well, input from serious participants in language-resource encoding (e.g. Nancy Ide, Adam Kilgarriff, Laurent Romary, …).

In one way or another, used in large-scale lexicon-oriented projects.

Ok, but how?

ISO LMF supplies a hand-crafted DTD with some example instances; there is **no consistent mapping procedure** formulated.

Find one. Look at what the XML community does.

Visual methods for modelling XML

Restriction: one visual modelling language: UML (mostly)

How standardized is XML modelling in UML?

Visual methods for modelling XML

Restriction: one visual modelling language: UML (mostly)

How standardized is XML modelling in UML?

GASP

It isn't.

Visual methods for modelling XML

Restriction: one visual modelling language: UML (mostly)

How standardized is XML modelling in UML?

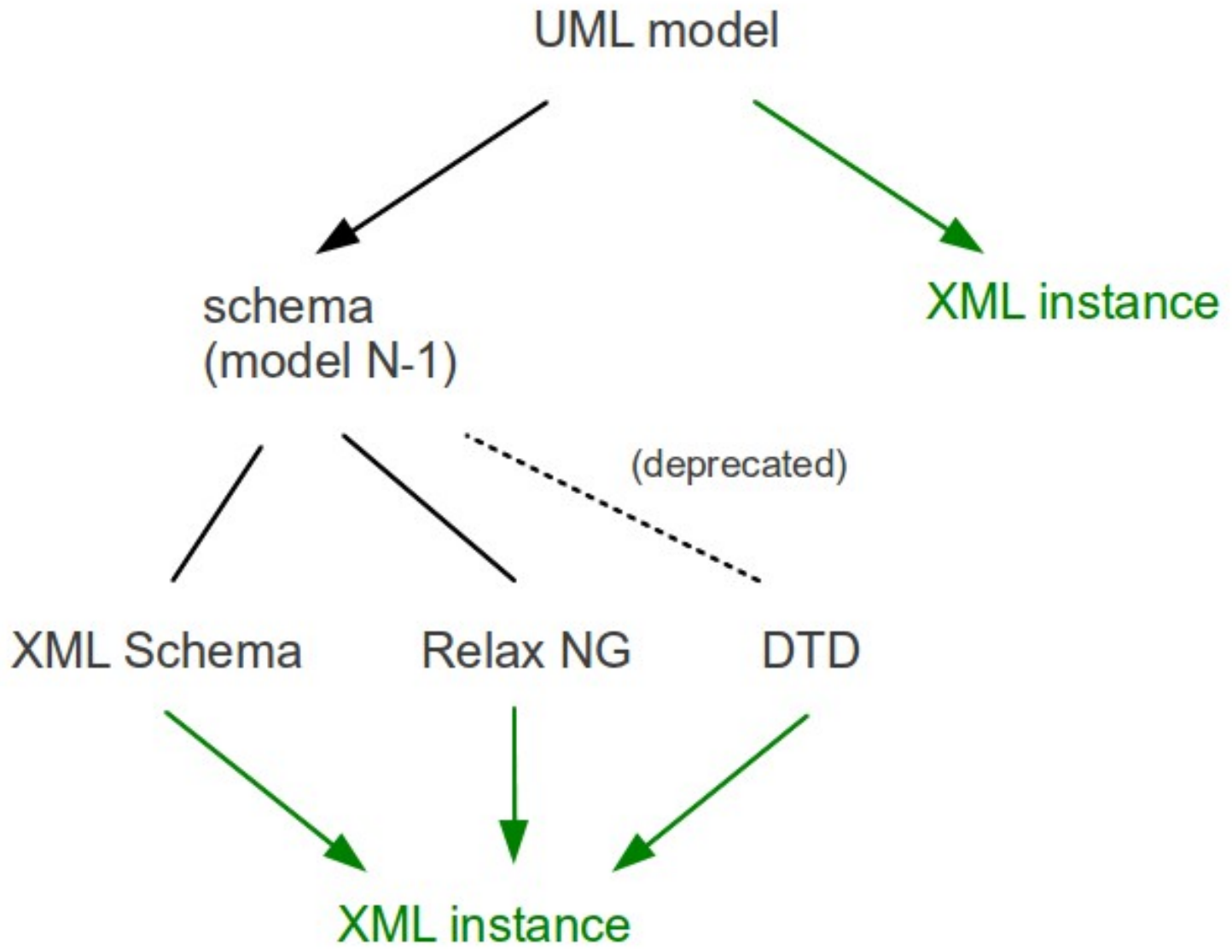
GASP

It isn't.

But an incredible amount of work has been done on this topic. Quite overwhelming, in fact.

Let us have a look at a very simple picture:

Simple picture of UML modelling of XML



Why XML Schema?

There's a dozen methods of going from UML to XML, and a majority of them, more or less directly, target XML Schema.

In the context of the TEI and my project, I tend to care about RNG a bit more. So why is it just XSD??

Speculations:

- standardised by W3C and, by virtue of that,
- more popular, head start with tool support (?),
- may be the choice of the data-oriented XML world over RELAX NG, which in turn is, possibly, more popular in document-oriented circles (or is it my TEI glasses?),
- targets vocabulary, so possibly it is more likely that one visually identifies a UML class with an XML element, and a UML class attribute with an XML attribute (But...)

On the other hand... why not RNG?

James Clark, http://www.thaiopensource.com/relaxng/design.html

“RELAX NG has the advantage in this role that it provides more flexibility in the choice of syntax.”

Eric van der Vlist, http://books.xmlschemata.org/relaxng/

“While there is a good overlap between UML and XML, the overlap isn't so good between XML and W3C XML schemas. (... [PSVI] … [On the other hand] ...) the overlap between UML, XML, and RELAX NG is almost as big as the overlap between UML and XML”

Question: if RNG is so good, why isn't it there?

RNG-users: total silence

XML-dev: Stephen D. Green on 15-06-2011:

- UML used to model OASIS TAML (Test Assertion Markup Language),
- via Martin Fowler's text syntax for UML

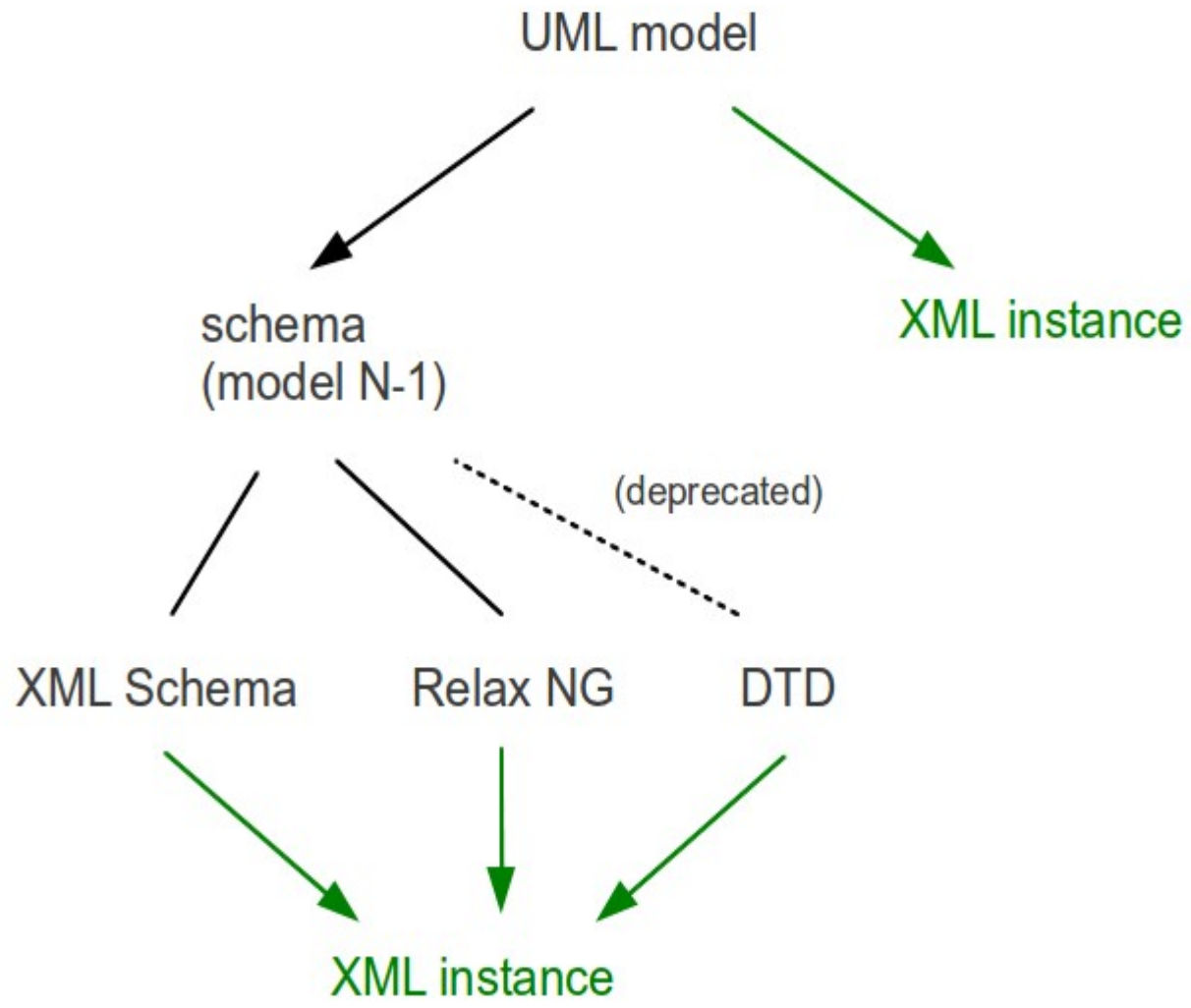
BTW, is it UML or MOF or EMF, or...?

No agreement on where exactly the top of the modelling hierarchy should be.

- Many say just “UML” (which UML? UML 2.0 is rather different, with different tools; the free ones for UML 1.x are mostly dead);
- Some say MOF (Meta-Object Facility), which is one level up wrt UML;
- Some say EMF (Eclipse Modeling Framework), a somewhat simpler equivalent of the MOF in the Eclipse coding universe.

Many modelling layers possible

- UML itself is defined by a metamodel, in MOF; some proposals use that;
- The UML layer can be split into the conceptual level and the logical level, i.e., a specialization of UML (UML profile); Dominguez et al. (2005): 3 layers.
- The UML-Schema transformation usually proceeds via an intermediate serialization, XMI (B. Marchal: XMI isn't fully standardised; Wikipedia: “Dysfunctional interchange format”)



So what about the TEI again?

Keep in mind my pet project: pairing UML with TEI

TEI is RELAX-NG-based, via its literate-encoding domain-specific language, ODD.

(Kudos to Michael Sperberg-McQueen, Syd Bauman, Lou Burnard, Sebastian Rahtz, Laurent Romary, and **many** others)

ODD is an overlay on RELAX NG that supplies the functionality needed for constructing and documenting TEI schemas. It doesn't need to rely on RELAX NG, but that was the (second) design choice for the TEI.

ODD has been used beyond the TEI:

- in ISO work (MLIF, MAF, TMF, TimeML) and
- in W3C practice: W3C ITS spec has been written as an ODD document.

Just to remind you...

From ch. 22 of the TEI Guidelines, a piece of ODD customisation:

```
<schemaSpec ident="example">
   <moduleRef key="teiheader"/>
   <moduleRef key="teistructure"/>
   <elementSpec ident="head" mode="change">
      <content>
         <rng:ref name="macro.xtext"/>
      </content>
   </elementSpec>
</schemaSpec>
```

What I am targeting is a single module, defined by the Dictionaries chapter (actually, a subpart of it).

(At this point, we should switch to oXygen for a moment).

A wrap-up and a suggestion

TEI:

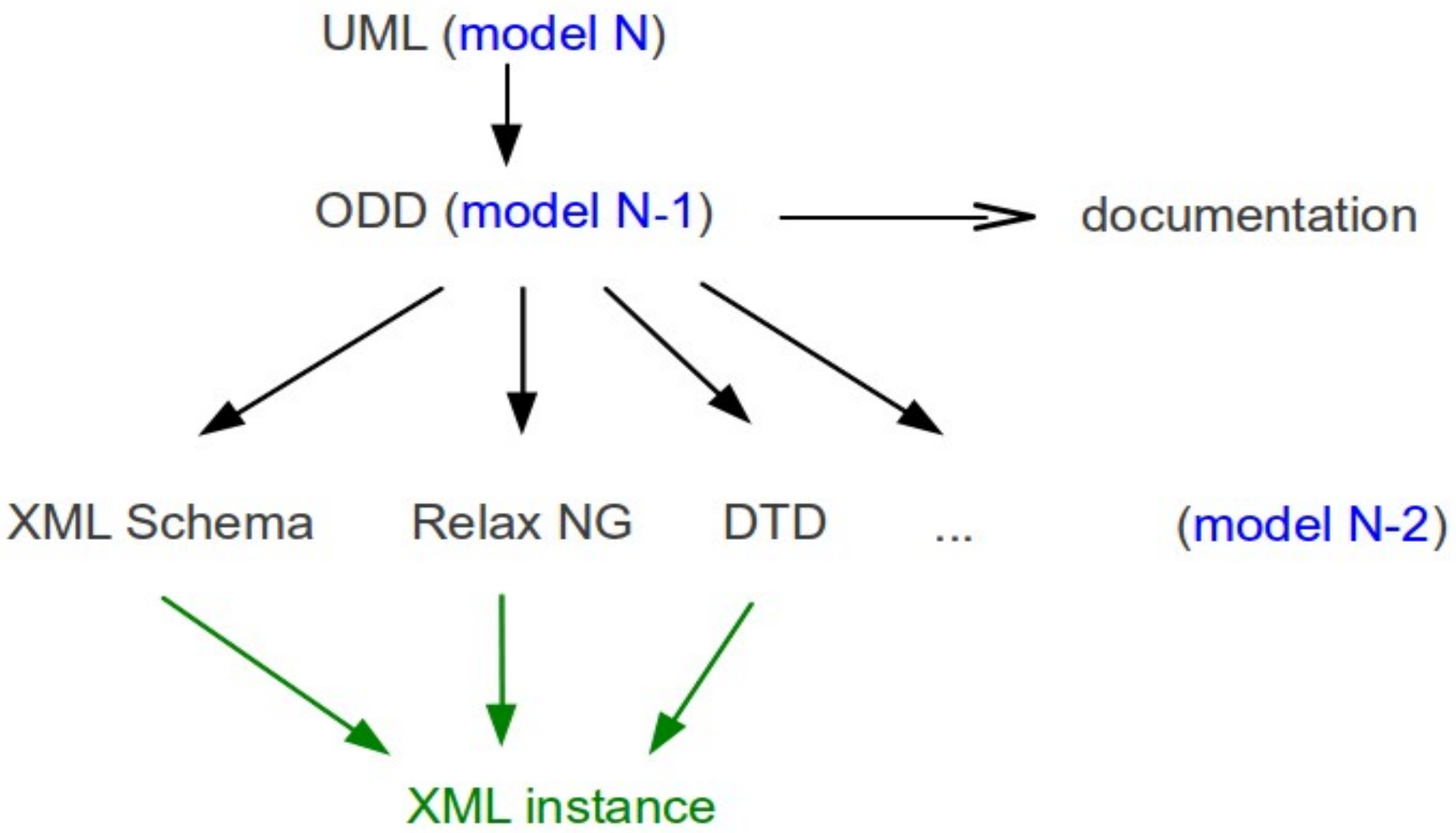
- parts of it are *de facto* standards in many areas of digital scholarship,
- advocated as the standard for linguistic markup in the CLARIN project,
- recommendations for dictionary markup are not a standard, although they deserve to become part of one.

ISO LMF is a standard almost without implementation. But it makes sense and shouldn't be buried yet!

There is no standard way of modelling schemas for XML.

Can these bits and pieces be wrapped together to create something new? I think it may be worthwhile to try.

Use ODD as the intermediate modelling layer



Conclusion...s

- ODD as one more level of indirection,
    - for those who'd rather not decide, and
    - for those who want to work with tighter models within the TEI;
- a ***literate*** level of indirection, mind you;
- it's nicely ***customisable***, too!
- possibly, just possibly, the data-oriented world might be interested in this;
- if something **useful** is called a standard but doesn't act as a standard (and Balisage has seen numerous instances thereof), then possibly, before it's buried, one way to get it off the ground may be to couple it with yet another, live standard. This is guaranteed not to work in most cases, BUT....

Thanks!

(and do share your thoughts, please)

bansp@o2.pl