



## Beyond DTDs: Constraining Data Content

José Carlos Ramalho  
Pedro Rangel Henriques

Language Processing and Specification Group  
Computer Science Department  
University of Minho  
Portugal




## Work Background



- **Project DAVID (1995) - “Algebraic Document Processing”**
- **Goal: Document Programming Environment**
- **Documents  $\cong$  Programs**
  - Both have to be processed
  - The steps towards processing are the same
- **Platform for format conversion**
- **Definition of a Doc. Program. Language**

## Project DAVID



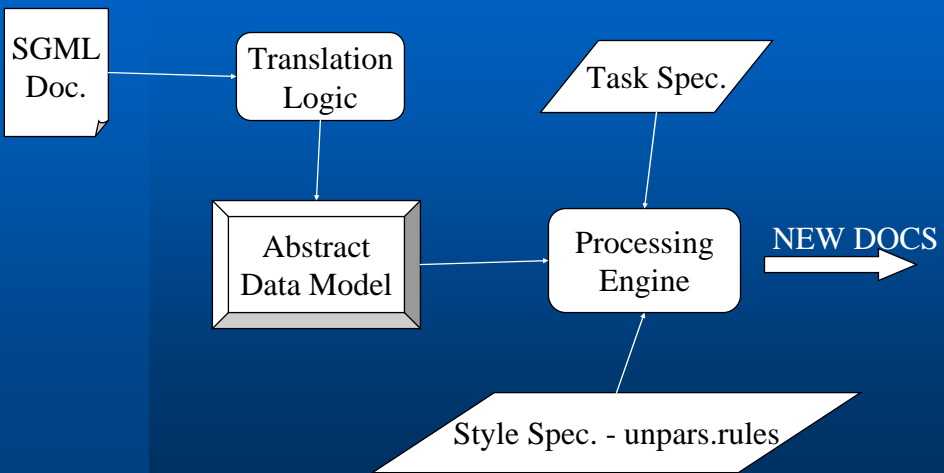
```

    graph LR
      Input[INPUT  
SGML] --> DPE[DPE]
      DPE --> Output[OUTPUT  
PS, HTML, RTF, ...]
  
```

- Just taking advantage of syntax
- Putting aside all the work done about parsing

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 3


## First Prototype



```


    graph LR
      SGML[SGML Doc.] --> TL[Translation Logic]
      TL --> ADM[Abstract Data Model]
      ADM --> PE[Processing Engine]
      TS[/Task Spec./] --> PE
      SS[/Style Spec. - unpars.rules/] --> PE
      PE --> NEW[NEW DOCS]
  
```

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 4



## Evolution


- Good ideas - too academic!?
- Lack of practice
- The Need to work with existing SGML models and Apps



To use this framework to process Semantics  
in  
SGML Documents

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho

5



## What will we discuss?

- Concerning SGML authoring, is Structural validation enough?
- Can we let the user to have full control of data?
- How can we Constrain more than Syntax?
- Is this problem a problem?
- What is its relevance?

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho

6

## What are we doing with SGML?



- **Constructing document DBs**
- **Publishing books on Internet and paper**
- **Converting parish registers (XIII and XIV century) to SGML**
- **Publishing from SGML DBs: Internet, CDROM, paper, ...**
- **Connecting SGML Documents to GIS**
- **Encoding Archives Documents**

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho

7

## Document Essence



- **What is a Document?**
- **What is its purpose?**
- **Who is its target audience?**
  
- **Must we have some special care regarding any of these items?**

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho

8

## Document is ...

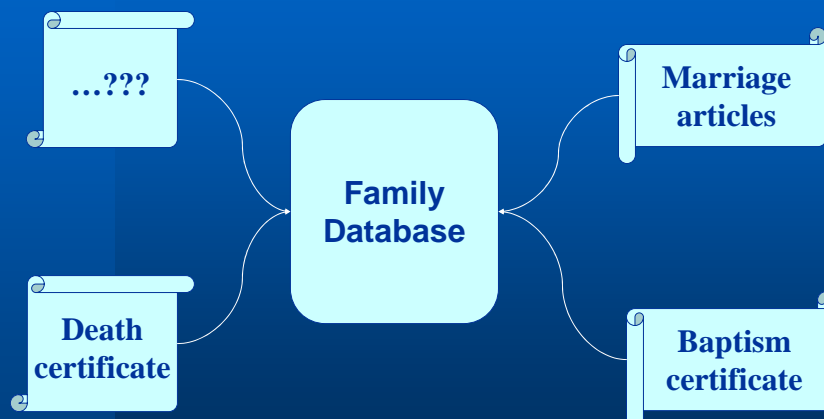


- “A paper with something written on” - 90% of the people
- “Something that teaches or can be used as an evidence” - from dic. encyc.
- “An information registry: a fact, an action, ...
  - its structure can range from very simple to very complex: a letter; a book
  - it can have multiple forms

## Existing SGML applications ...



### Case Study 1: Parish register (XIII and XIV century)



## Existing SGML applications ...

Case Study 1: Parish register (XIII and XIV century)

**Problems:**

- negative ages
- death before baptism
- marriages between people with age differences higher than 100
- ...

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 11

## Existing SGML applications ...

Case Study 2: Archaeology (Sites and Artefacts)

Arch.Site record

Arch.Artefact record

SGML database

Internet (www)

GIS

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 12

## Existing SGML applications ...

### Case Study 2: Archaeology (Places and Artefacts)

Arch. Place record

Arch. Artefact record

SGML database

Internet (www)

Every coordinate (latitude and longitude) falls into the map in question

GIS

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho

13

## SGML and Semantics

- Can we just add constraints to SGML specifications in order to process semantic validations?
- What is missing?
- Do we need more than just constraints?

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho

14

## SGML: What ...



- **SGML was conceived to specify structure**
- **Today, it is a very powerful specification tool**
- **We can not use its syntax to express constraints or invariants over element content**

## SGML: How ...



- **To add extra markup**
- **To design a complete new language that could be embedded in SGML or coexist outside. Ex: database applications embed SQL**



## The Constraints



**90% of the cases are very simple**

- **To restrict atomic element values**
- **To check relations between elements**
- **To perform a lookup operation with a value in some database**

We feel the simplicity derives from the fact that SGML takes care of validations at higher levels (the structural ones).

## The Semantic Validation Model



**2 different steps:**

- **the definition: the syntactic part; the statements that express the constraints.**
- **the processing: the semantic part; the constraints interpretation.**

## The 2 Steps

- **The definition implies:**
  - the creation of a new language
  - or
  - the adoption of an existent one
- **The processing implies:**
  - the creation of an engine capable of processing constraints written in the language above

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 19

## Building a Semantic Val. Model

Definition Stage

Constraints

Processing Stage


SGML Document

?

Text?  
Structured Text?  
Typed Information?

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 20

## What about types?



**SGML Document**

...

```
<latitude>41.32</latitude>
```

...

**Constraint**


```
latitude > 39 and latitude < 43
```

- We are comparing an element against numeric values
- Those values have an inherent type (integer or float)
- The processing engine needs somehow to associate that type with the element being compared

Numeric types are easy to infer. What about other types?

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 21

## What about other types?



**SGML Document**

...

```
<date>19 May 1998</date>
```

...

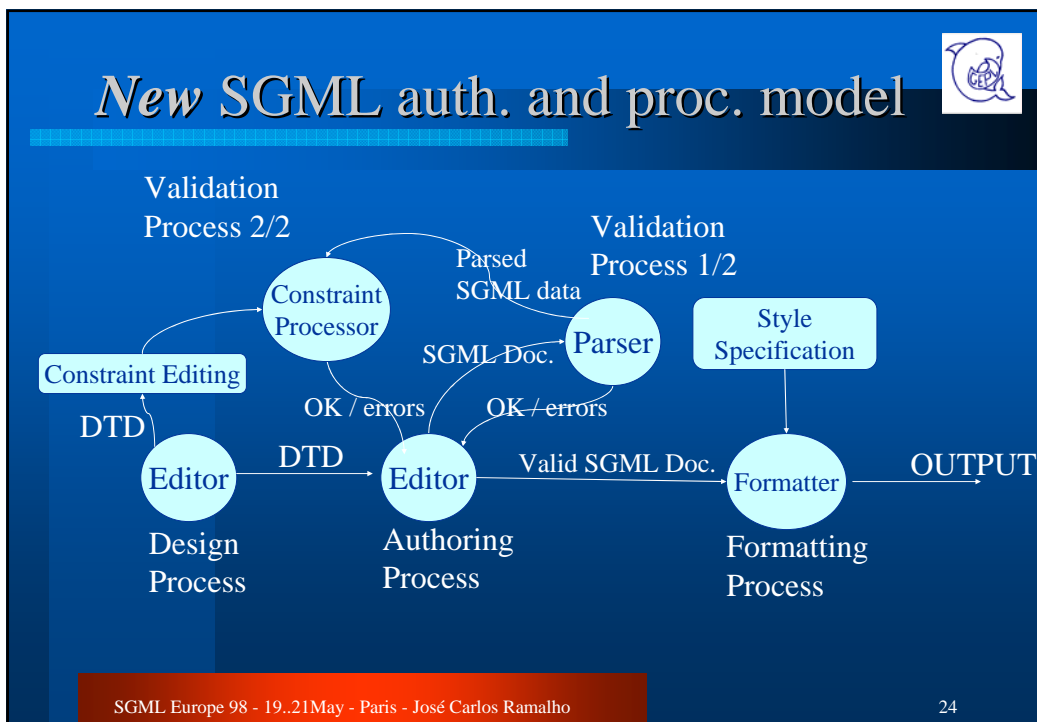
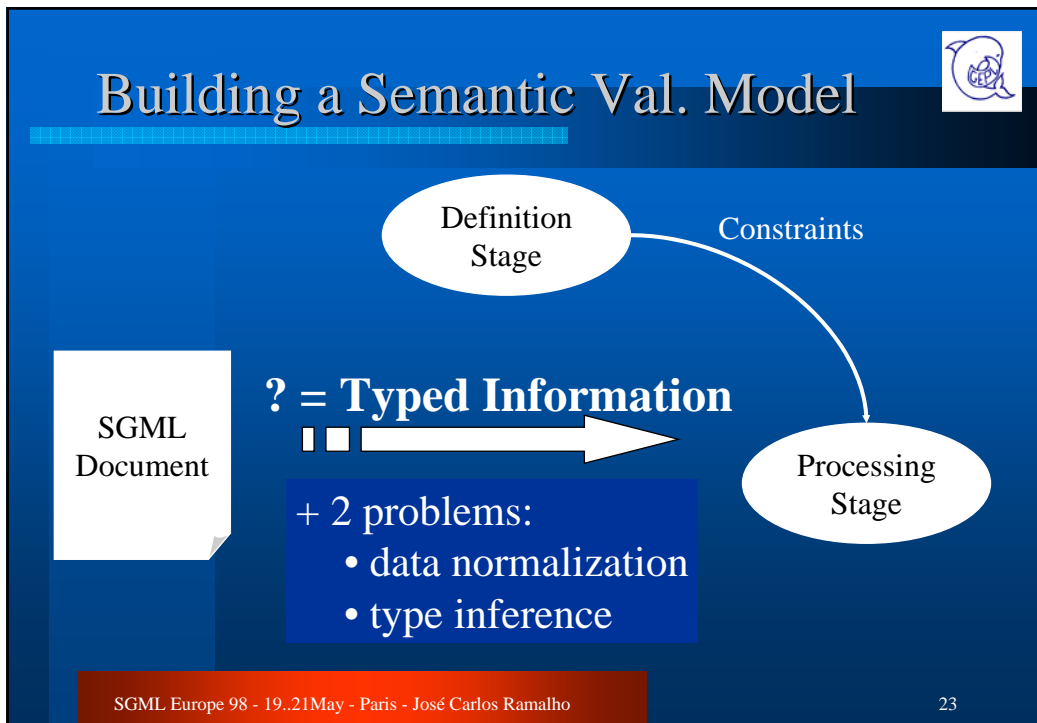
**Constraint**

```
date > 1998.05.15
```

**But we still have more types like lists ...**

- There are more than 100 date formats
- Here we have another problem:
  - Data Normalization

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 22



## Solutions



- **Developing complex tools to do the job:**
  - first the data normalization
  - then the type inference
- **Adding some extra definitions into the DTD and some extra markup to the SGML data**

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho

25

## The Solution



- **To define 2 extra attributes optional to all elements:**
  - **value** for data normalization (as in TEI DTD)
  - **type** for type inference
- **Examples:**

... it happened in `<date type="date" value="1853.10.05">` the fifth of October of the year 1853 `</date>` ...

... `<latitude type="float">41.32</latitude>` ...

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho

26

## Do we need to type every element?

**Archaeological sites Doc. Structure**

```

graph TD
    arqsits --> arqelem
    arqelem --> identi
    arqelem --> latitu
    arqelem --> quadro
    arqelem --> interp
    arqelem --> dots[...]
    quadro --> liga
    quadro --> texto
  
```

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 27

## Do we need to type every element?

**A “no” answer implies:**

- simplicity in the constraining language
- simplicity in the processing engine implementation
- an incomplete abstract model;
  - **this would disable any manipulation of the document as an abstract model**

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 28

## Do we need to type every element?

### Archaeological sites Doc. Structure

**identi**

Type = "string"

**latitu**

Type = "float"

Type = "string"

**liga**

**texto**

Type = "text"

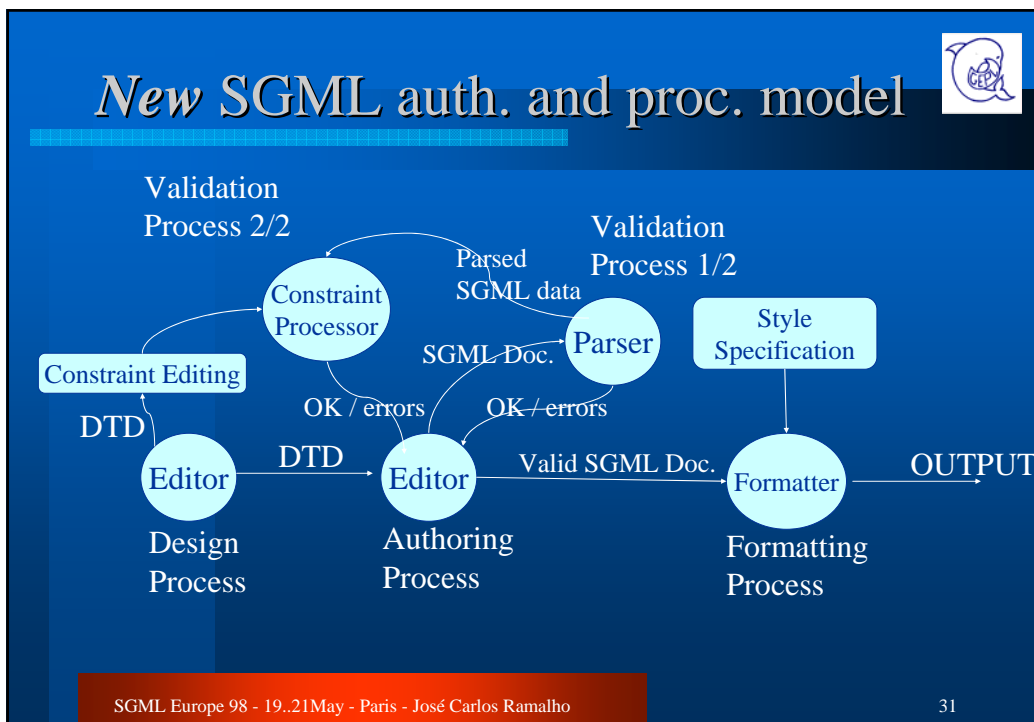
SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 29

## Do we need to type every element?

A "yes" answer would imply:

- a more complicated type system and processing engine
- probably the structured types are best inferred from the DTD
- **we are able to process the document as an abstract model**

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 30



## Are we reinventing the Wheel?

Looking for similarities ...

### SGML Documents and Programs

<ul style="list-style-type: none"> <li>● Program</li> <li>● Program. Language</li> <li>● Grammar Rules</li> <li>● Terminals: chars and words (grammar)</li> </ul>	<ul style="list-style-type: none"> <li>● SGML Document</li> <li>● Markup Language</li> <li>● DTD</li> <li>● Terminals: chars and words (SGML declaration and DTD)</li> </ul>
---	--

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 32



# Programming Languages Processing

- Lexical Analysis
- Syntactic Analysis
- Semantic Analysis
- Code Generation

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 33


# PL Processing Models

## Syntax Directed Translation (SDT)

- Lexical Anal. - specified and automatic
- Syntactic Anal. - specified and automatic
- Semantic A. - programmed by the developer

**that is the scenery we have with SGML processing model**

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 34




## PL Processing Models

### Semantics Directed Translation (based in Attribute Grammars - Knuth1968)

- Lexical Analysis - ...
- Syntactic Analysis - ...
- Semantic A. - **specified and automatic**

**This is our goal in SGML context**


SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 35



## Conclusion

- We have specified a way to add Semantics to SGML Documents
- We have identified the problems: data normalization and type inference
- We did not specify a Constraining Language but we know that it is not a problem
- The problem will be the Constraint Processing

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 36



## Future Work

- A simple constraint language is being studied/created.
- We are going to implement this semantic validation scheme (with the new language) in our prototype INES (“A Document Programming Environment”).

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 37



## Questions ...?

After Conference Contacts:

- Web: [www.di.uminho.pt/~jcr](http://www.di.uminho.pt/~jcr)
- Email: [jcr@di.uminho.pt](mailto:jcr@di.uminho.pt)

SGML Europe 98 - 19..21May - Paris - José Carlos Ramalho 38