

DON'T TOUCH MY SGML

SGML

Sounds Good. Maybe Later.

ari.nordstrom@gmail.com



CREATIVE WORDS

I LOVE SGML



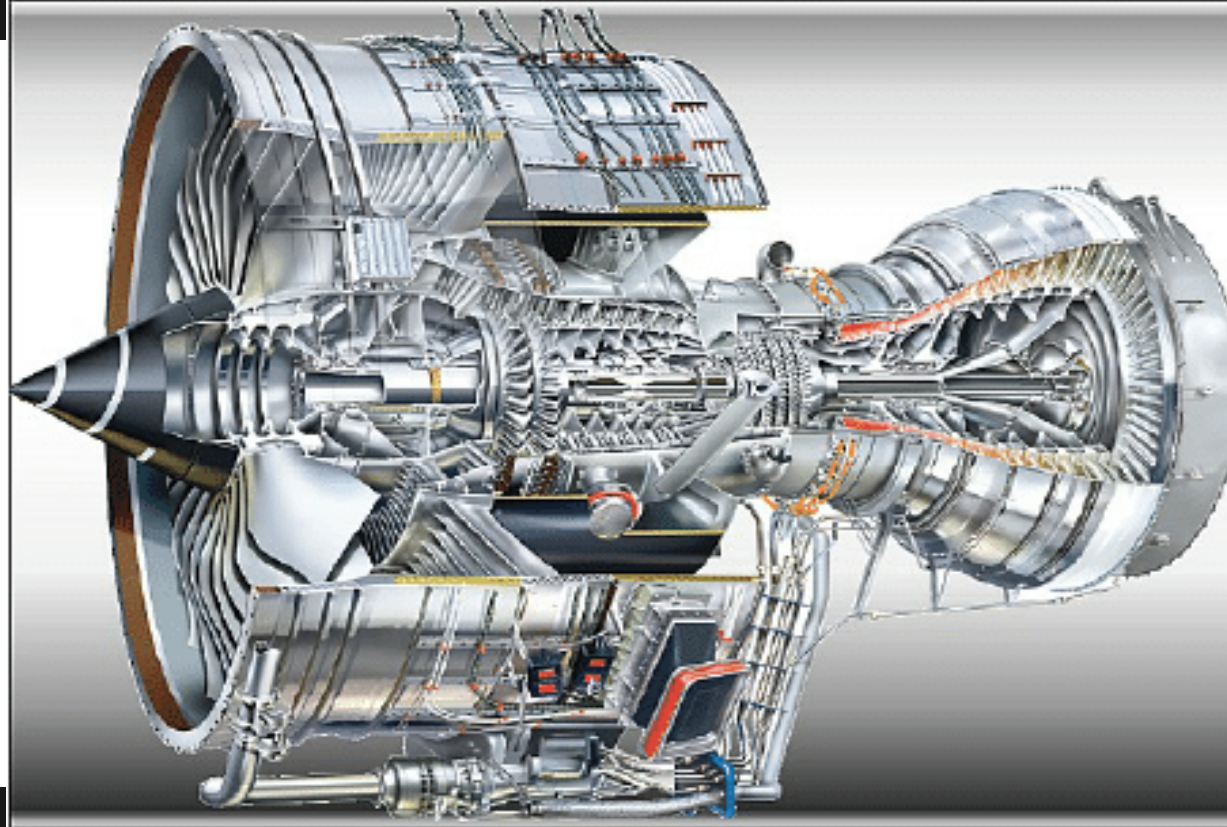
CREATIVE WORDS

BUT THE TIME WAS NEVER REALLY RIGHT



CREATIVE WORDS

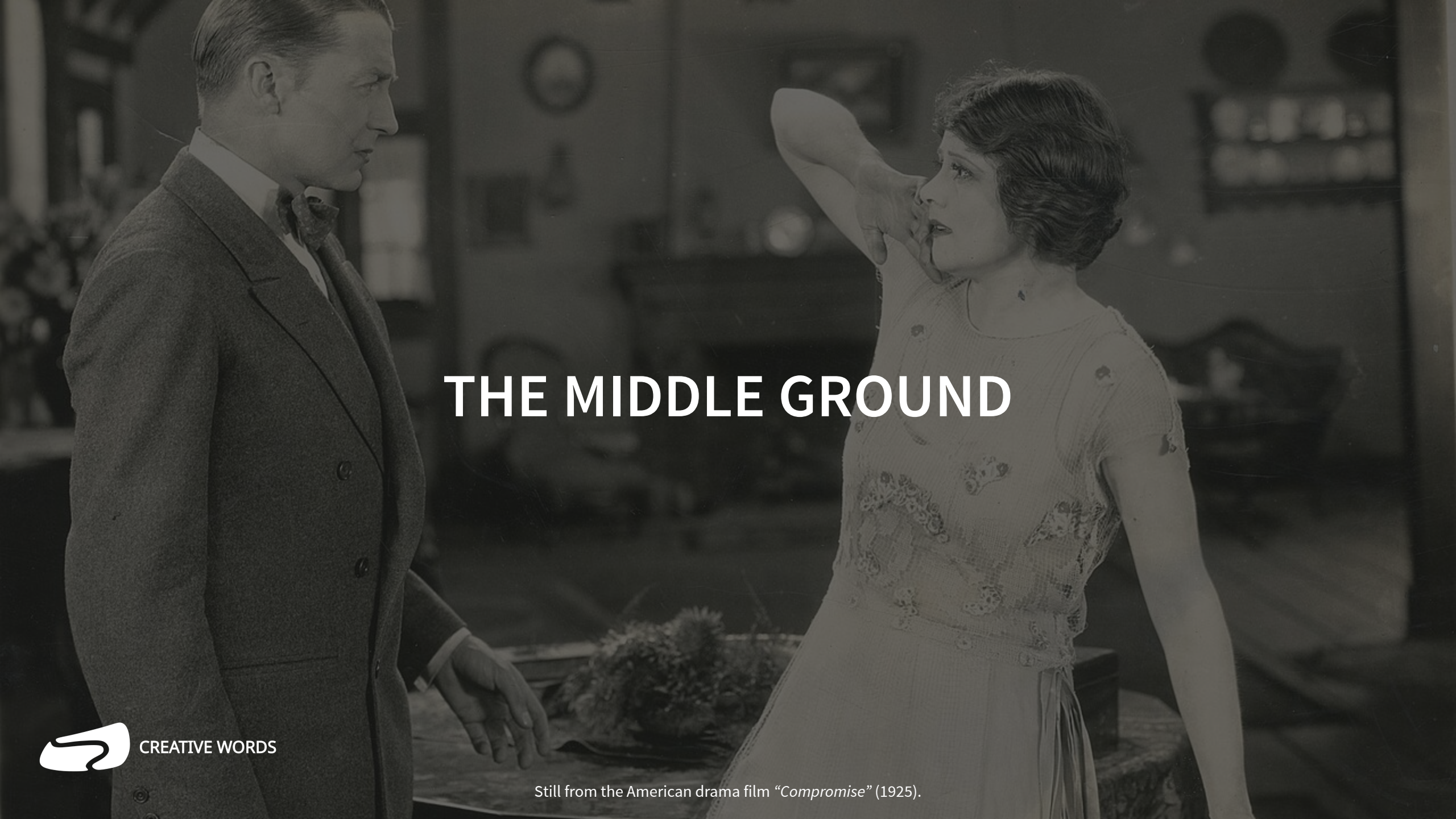
USEFUL AND SMART, THOUGH



ATA iSpec 2200 MIL-STD-38784 S1000D Issues 1.8, 1.9, 2.0

TODAY: SGML POORLY SUPPORTED

- There is a middle ground



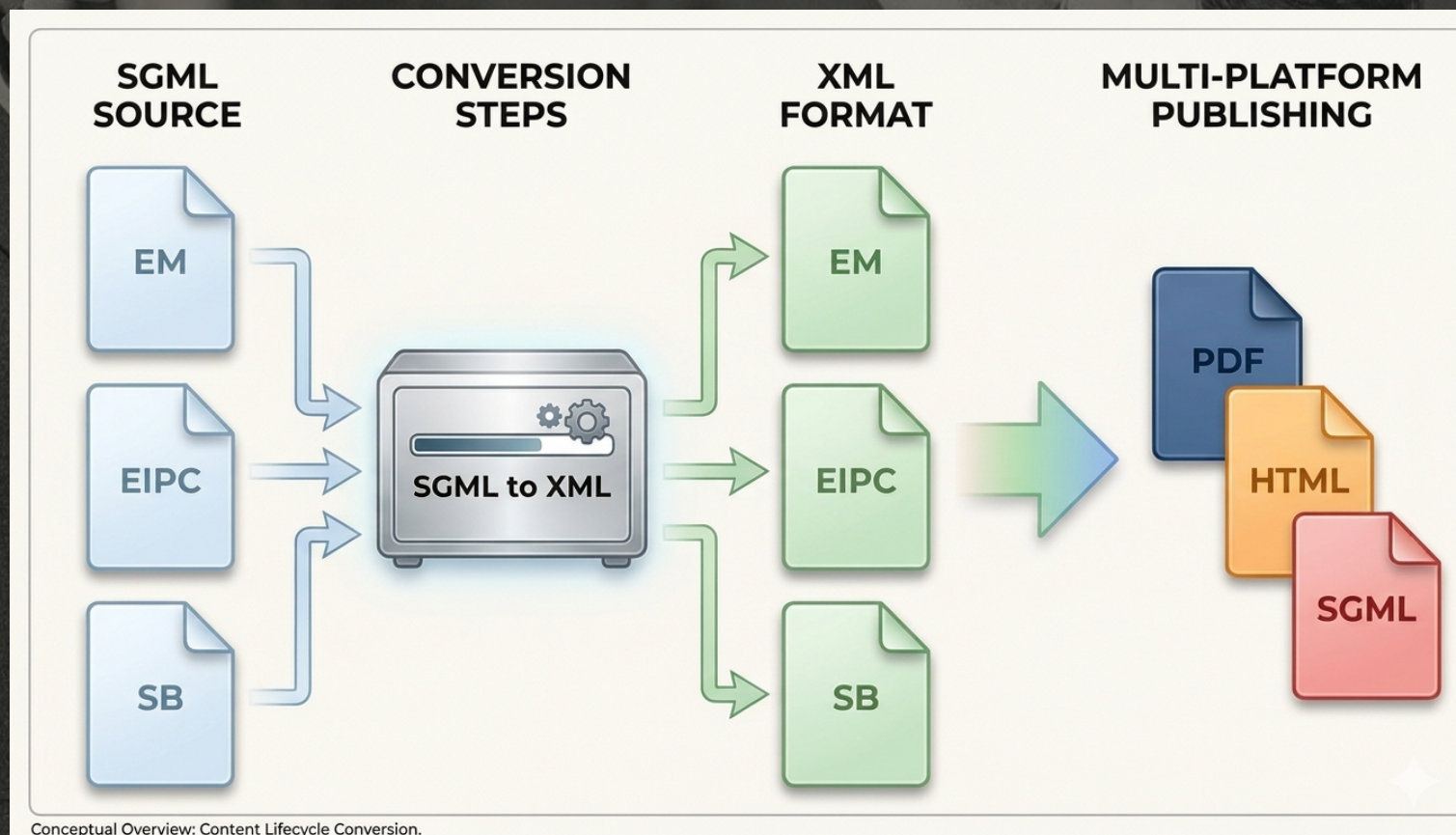
THE MIDDLE GROUND



CREATIVE WORDS

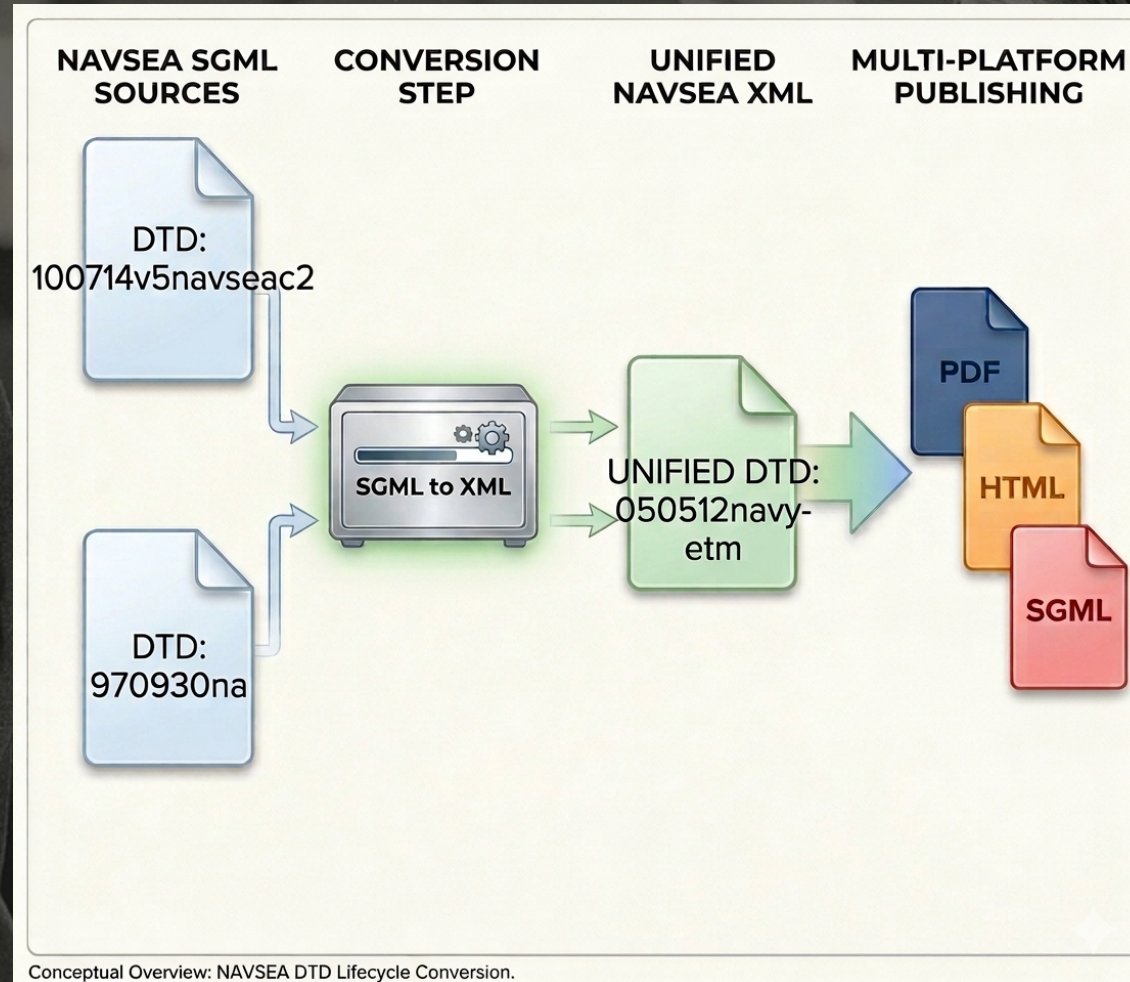
Still from the American drama film "*Compromise*" (1925).

ATA ISPEC 2200 SGML TO XML

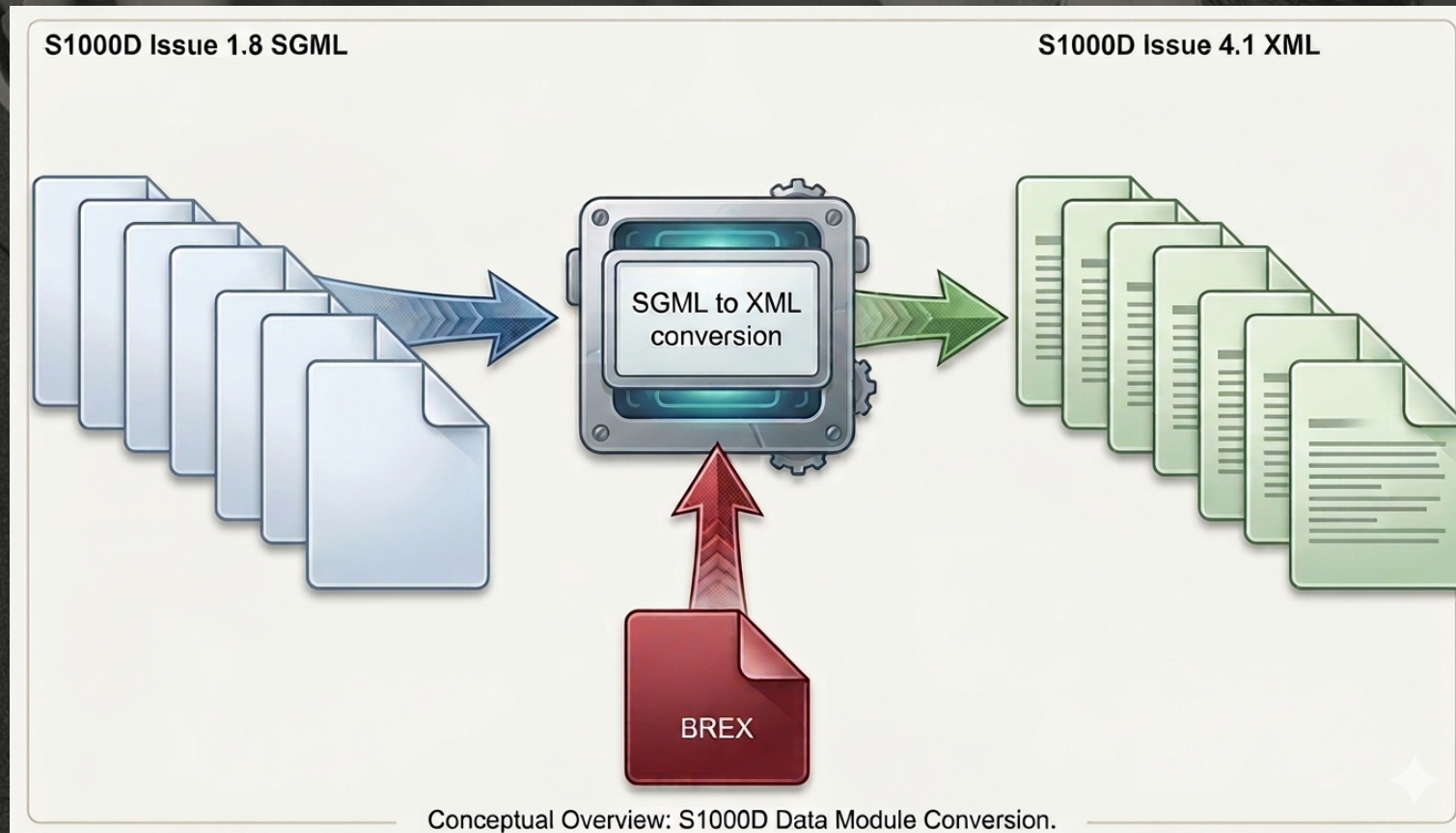


CREATIVE WORDS

MIL-STD-38784 SGML TO XML



S1000D 1.8 SGML TO 4.1 XML



Compromise

A poignant drama of today

IRENE RICH, the wife who fought a fair fight for love.

Clive Brook, who chose between the wife who waited to be wooed, and the flapper who did the wooing.

Pauline Garon, as the love thief.

Louise Fazenda, in the greatest comedy role of her career.

Raymond McKee as the



"I TOLD YOU I WOULD TAKE HIM FROM YOU"



CREATIVE WORDS

Poster for the film "Compromise".

NO SUCH THING AS WELL-FORMED SGML

You **must** have a DTD

Tag omission

Entities

Encoding (SDATA)

Syntax

(Converting an SGML DTD to XML is sometimes a ****)

Compromise

STRUCTURAL CHALLENGES

Inclusions

```
1 <!ELEMENT em - - ((title, legalntc?, mfmatr, chapter*)  
2 | increv | tr+)  
3 +(revst | revend | cocst | cocend | hotlink) >
```

But also

- Exclusions
- Logical AND groups
- Data types

• ...

Irene Rich as the wife who fought a fair fight.

Clive Brook, who chose between the wife who wooed, and the flapper who did the wooing.

Pauline Campbell, the love thief.

Louise Fazenda, in the greatest comedy role of her career.

Raymond McKee as the

" I TOLD YOU I WOULD TAKE HIM FROM YOU

Compromise

SGML DECLARATION

All kinds of fun features, like changing tag delimiters

```
{doc}  
  {para}Some text.  
  {para}More text.  
{=doc}
```

This is where I drew the line.

Pauline Garon, as the
love thief.

Louise Fazenda, in the
greatest comedy role of her career.

Raymond McKee as the

" I TOLD YOU I WOULD TAKE HIM FROM YOU

Compromise

TOOLS?

James Clark's SP package (aka OpenSP)

Appointed drama of today
IRENE RICH as the
wife who fought

nsgmls (parser)

Clive Brook as the
between the wife who waited to be
wooed, and the flapper who did the
wooing.

spam (normaliser)

Pauline Garon, as the
love thief.

sx (XML syntax converter)

Louise Fazenda, in the
greatest comedy role of her career.

Raymond McKee as the

" I TOLD YOU I WOULD TAKE HIM FROM YOU



CREATIVE WORDS

AN SGML DTD...

```
1 <!-- (NOTATION declarations) -->
2
3 <!ENTITY % id.att      "id ID #IMPLIED" >
4
5 <!ELEMENT doc          - - (title, chapter+) >
6 <!ATTLIST doc          %id.att; >
7
8 <!ELEMENT title        - - (#PCDATA) >
9
10 <!ELEMENT chapter      - - (title, (para | graphic)+, section*) >
11 <!ATTLIST chapter      %id.att; >
12
13 <!ELEMENT section      - - (title, (para | graphic)+, section*) >
14 <!ATTLIST section      %id.att; >
15
16 <!ELEMENT para         - - (#PCDATA) >
17
18 <!ELEMENT graphic      - 0 EMPTY >
19 <!ATTLIST graphic      href CDATA #IMPLIED
20                        name ENTITY #IMPLIED
21                        %id.att; >
```

...AND A VALID SGML INSTANCE

```
1 <!DOCTYPE doc PUBLIC "-//CWORD//DTD DOC TEST//EN" "doc.dtd" [  
2 <!ENTITY bill SYSTEM "Bill_Murray.jpg" NDATA jpeg >  
3 ]>  
4 <doc>  
5     <title>My Test SGML Document</title>  
6     <chapter>  
7         <title>Chapter 1 Title</title>  
8         <para>Some content.</para>  
9         <graphic href="Bill_Murray.jpg">  
10    </chapter>  
11    <chapter>  
12        <title>Chapter 2 Title</title>  
13        <para>More content.</para>  
14        <section>  
15            <title>Subsection 1 Title</title>  
16            <para>Subsection content.</para>  
17            <graphic name="bill">  
18        </section>  
19    </chapter>  
20 </doc>
```

BUT NOT EVERYTHING IS UNICODE

```
<!ENTITY % ISOtech PUBLIC
"ISO 8879-1986//ENTITIES General Technical//EN"
>
<!ENTITY % ISOpub PUBLIC
"ISO 8879-1986//ENTITIES Publishing//EN"
>
<!ENTITY % ISOnum PUBLIC
"ISO 8879-1986//ENTITIES Numeric and Special Graphic//EN"
>
<!ENTITY % ISOgrk1 PUBLIC
"ISO 8879-1986//ENTITIES Greek Letters//EN"
>
<!ENTITY % ISOgrk2 PUBLIC
"ISO 8879-1986//ENTITIES Monotoniko Greek//EN"
>
<!ENTITY % ISOgrk3 PUBLIC
"ISO 8879-1986//ENTITIES Greek Symbols//EN"
>
<!ENTITY % ISOgrk4 PUBLIC
"ISO 8879-1986//ENTITIES Alternative Greek Symbols//EN"
>
<!ENTITY % ISOlat1 PUBLIC
"ISO 8879-1986//ENTITIES Added Latin 1//EN"
>
<!ENTITY % ISObox PUBLIC
"ISO 8879-1986//ENTITIES Box and Line Drawing//EN"
>
```

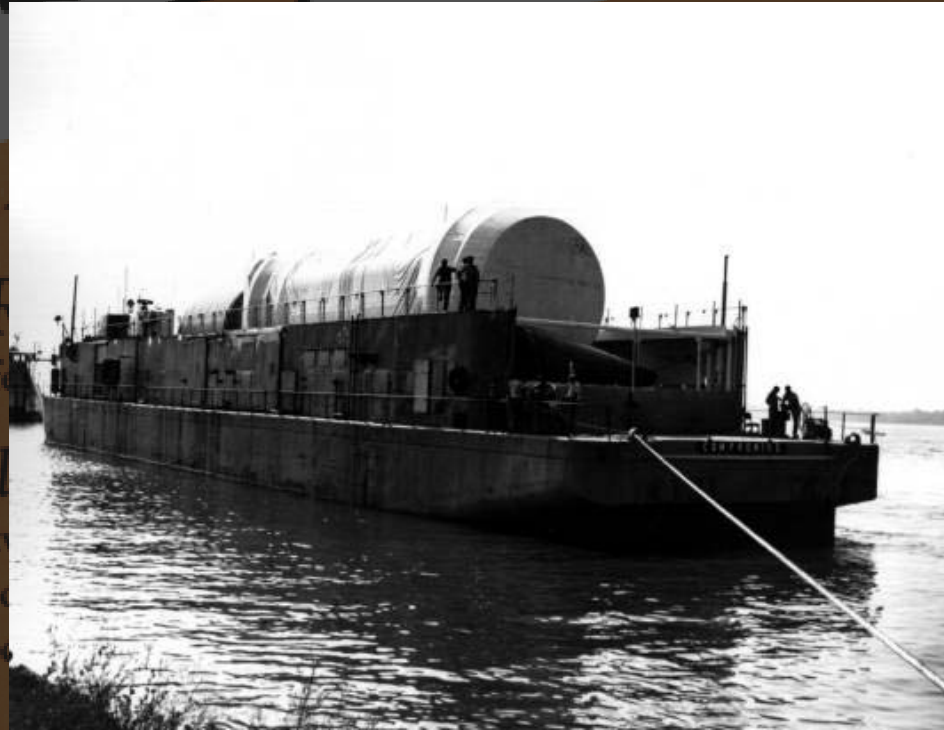
```
%ISOtech; %ISOpub; %ISOnum; %ISOgrk1; %ISOgrk2; %ISOgrk3;
%ISOgrk4; %ISOlat1; %ISObox;
```

CHARACTER ENTITIES ARE A PAIN

—

```
<!ENTITY mdash SDATA "[mdash]"--=em dash-->
```

WE'LL NEED A FEW STEPS



"Compromise" barge carrying Saturn V rocket sections

Louise Fazenda, in the greatest comedy role of her career.

Raymond McKee as the

"I TOLD YOU I WOULD TAKE HIM FROM YOU"

NORMALISATION WITH spam

```
<!DOCTYPE doc [  
<!ENTITY bill SYSTEM "Bill_Murray.jpg" NDATA jpeg >  
  
<!-- NOTATIONS -->  
  
<!ENTITY % id.att "id ID #IMPLIED" >  
  
<!ELEMENT doc - - (title, chapter+) >  
<!ATTLIST doc %id.att; >  
  
<!ELEMENT title - - (#PCDATA) >  
  
<!ELEMENT chapter - - (title, (para | graphic)+, section*) >  
<!ATTLIST chapter %id.att; >  
  
<!ELEMENT section - - (title, (para | graphic)+, section*) >  
<!ATTLIST section %id.att; >  
  
<!ELEMENT para - - (#PCDATA) >  
  
<!ELEMENT graphic - 0 EMPTY >  
<!ATTLIST graphic href CDATA #IMPLIED  
name ENTITY #IMPLIED  
%id.att; >]>
```

XML SYNTAX WITH SX

...PLUS A SET OF CHARACTER ENTITIES

```
1 <?xml version="1.0"?>
2 <!DOCTYPE doc [<!-- This maps SGML character entities to their Unicode equivalents -->
3 <!-- Based on DocBook 4.1.2 XML entities --><!-- ISO 8879 official entity sets -->
4 <!ENTITY % iso-amsa PUBLIC "ISO 8879:1986//ENTITIES Added Math Symbols: Arrow Relatio
5 %iso-amsa;
6 <!ENTITY % iso-amsb PUBLIC "ISO 8879:1986//ENTITIES Added Math Symbols: Binary Operat
7 %iso-amsb;
8 <!ENTITY % iso-amsc PUBLIC "ISO 8879:1986//ENTITIES Added Math Symbols: Delimiters//E
9 %iso-amsc;
10 <!ENTITY % iso-amsn PUBLIC "ISO 8879:1986//ENTITIES Added Math Symbols: Negated Relat
11 %iso-amsn;
12 <!ENTITY % iso-amso PUBLIC "ISO 8879:1986//ENTITIES Added Math Symbols: Ordinary//EN"
13 %iso-amso;
14 <!ENTITY % iso-amsr PUBLIC "ISO 8879:1986//ENTITIES Added Math Symbols: Relations//EN
15 %iso-amsr;
16 <!ENTITY % iso-lat1 PUBLIC "ISO 8879:1986//ENTITIES Added Latin 1//EN" "/home/ari/Doc
17 %iso-lat1;
18 <!ENTITY % iso-lat2 PUBLIC "ISO 8879:1986//ENTITIES Added Latin 2//EN" "/home/ari/Doc
19 %iso-lat2;
20 <!ENTITY % iso-grk4 PUBLIC "ISO 8879:1986//ENTITIES Alternative Greek Symbols//EN" "/"
21 %iso-grk4;
```

XML-BASED PROCESSING

```
1 <manifest xmlns="http://www.corbas.co.uk/ns/transforms/manifest" xml:base=".">
2
3   <group description="Main structures" xml:base="../../xslt/ata-xml/">
4     <item href="ATAXML_table.xsl" description="Convert SGML CALS to OASIS CALS"/>
5     <item href="ATAXML_ulink.xsl" description="Convert ulink to refext"/>
6   </group>
7
8   <group description="EM-specific steps" xml:base="../../xslt/ata-em/">
9     <item href="lXitem-effect.xsl" description="Move effect outside lXitem if first" />
10    <item href="EM_nested-elem2pi.xsl" description="Handle nested inclusion elements" />
11    <item href="EM_prc-elem2pi.xsl" description="Convert inclusion elements in proc" />
12  </group>
13
14  <group description="Common effect handling" xml:base="../../xslt/ata-xml/">
15    <item href="ATAXML_move-effect.xsl" description="Move misplaced effect element" />
16    <item href="ATAXML_elem2pi.xsl" description="Handle non-nested inclusion elements" />
17  </group>
18
19  <group description="Common other" xml:base="../../xslt/common/">
20    <item href="empty-sbdata.xsl" description="Fix empty SB list for EM and SDS"/>
21  </group>
```

ORCHESTRATED IN ANT

```
<exec
  executable="${sp-loc.path.spam}"
  output="${local.spam}\${local.filename}"
  dir="."
  inputstring="#10;">
  <arg value="-p"/>
  <arg value="-p"/>
  <arg value="-x"/>
  <arg value="-x"/>
  <arg value="-l"/>
  <arg value="-D${base.source}"/>
  <arg value="-E0"/>
  <arg value="-ferrspam.txt"/>
  <arg value="-c"/>
  <arg value="${module.schemas.input}"/>
  <arg value="${file}"/>
</exec>
```

Compromise

ATA, MIL-SPEC, S1000D...

- Most steps are the same
- Differences on the XML side

A poignant drama of today
IRENE RICH as the
wife who waited for her lover

Click, Break... the story
between the wife who waited to be
wooed, and the flapper who did the
wooing.

Pauline Garon, as the
love thief.

Louise Fazenda, in the
greatest comedy role of her career.

Raymond McKee as the

" I TOLD YOU I WOULD TAKE HIM FROM YOU

Compromise

A poignant drama of today

IRENE RICH as the

Is there a way to standardise all this?

Clive Brook, who chose between the wife who waited to be wooed, and the flapper who did the wooing.

Pauline Garon, as the love thief.

Louise Fazenda, in the greatest comedy role of her career.

Raymond McKee as the

"I TOLD YOU I WOULD TAKE HIM FROM YOU"



CREATIVE WORDS

SGML2XML-TNG

 **sgml2xml**

ant

catalogs

lib ◦

modules ◦

build.properties.xml

build.xml



CREATIVE WORDS

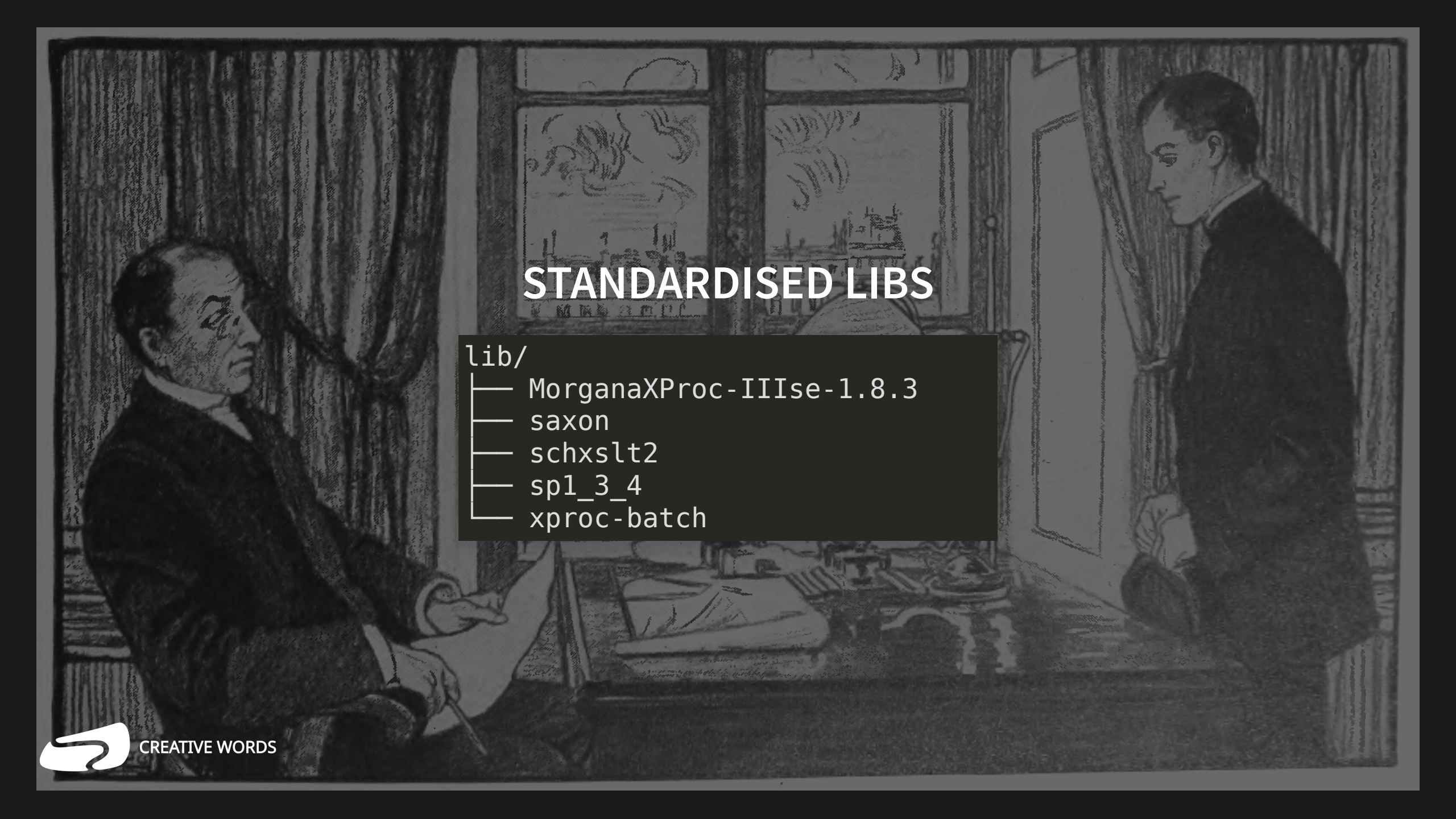
Illustration on p. 203 , Harper's Magazine Vol 133, 1916--story "An Incidence in the Prefecture of Police" by Arthur Sherburne Hardy; illustrator, Elizabeth Shippen Green. [Caption: IT APPEARS TO ME YOU WISH ME TO COMPROMISE WITH DUTY]

A black and white illustration of a man sitting at a desk and another man standing by a window, with the word 'STILLANT' overlaid. The man on the left is seated at a desk, looking towards the right. He is wearing a dark suit and a white shirt. The desk in front of him is cluttered with papers, a pen, and other writing instruments. To his right, a large window looks out onto a cityscape with buildings and a bridge. A lamp is positioned on the desk near the window. On the far right, another man stands by the window, looking out. He is also wearing a dark suit and a white shirt. The word 'STILLANT' is written in large, white, capital letters across the center of the image, partially obscuring the window and the desk. The overall style is that of a classic illustration, possibly a woodcut or a detailed drawing.

STILLANT



CREATIVE WORDS



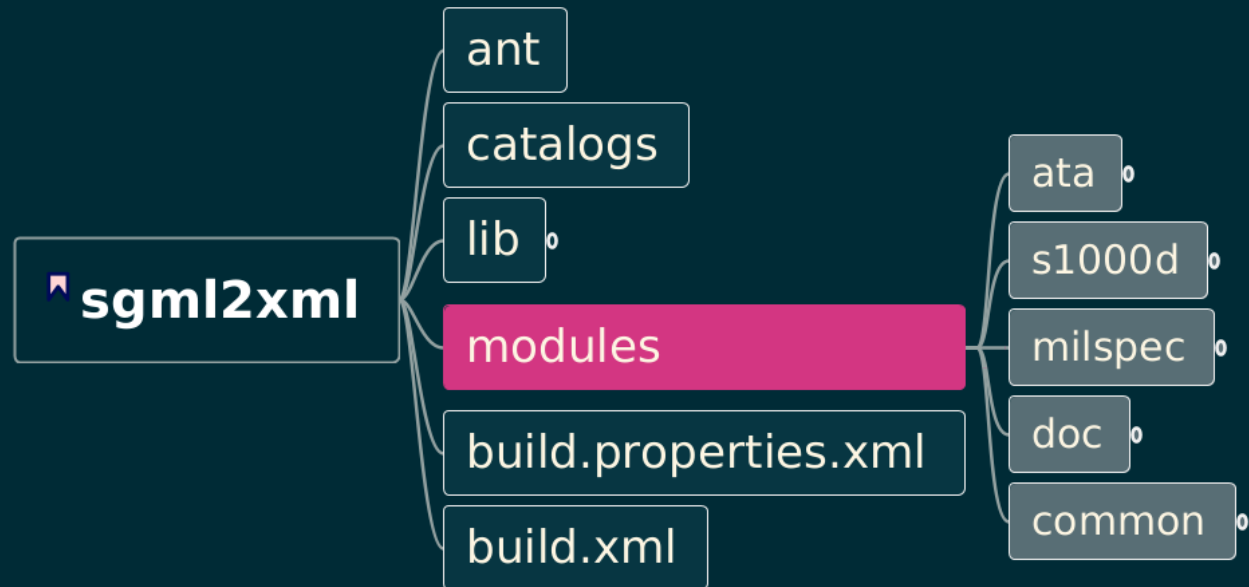
STANDARDISED LIBS

```
lib/  
├── MorganaXProc-IIIse-1.8.3  
├── saxon  
├── schxslt2  
├── sp1_3_4  
└── xproc-batch
```

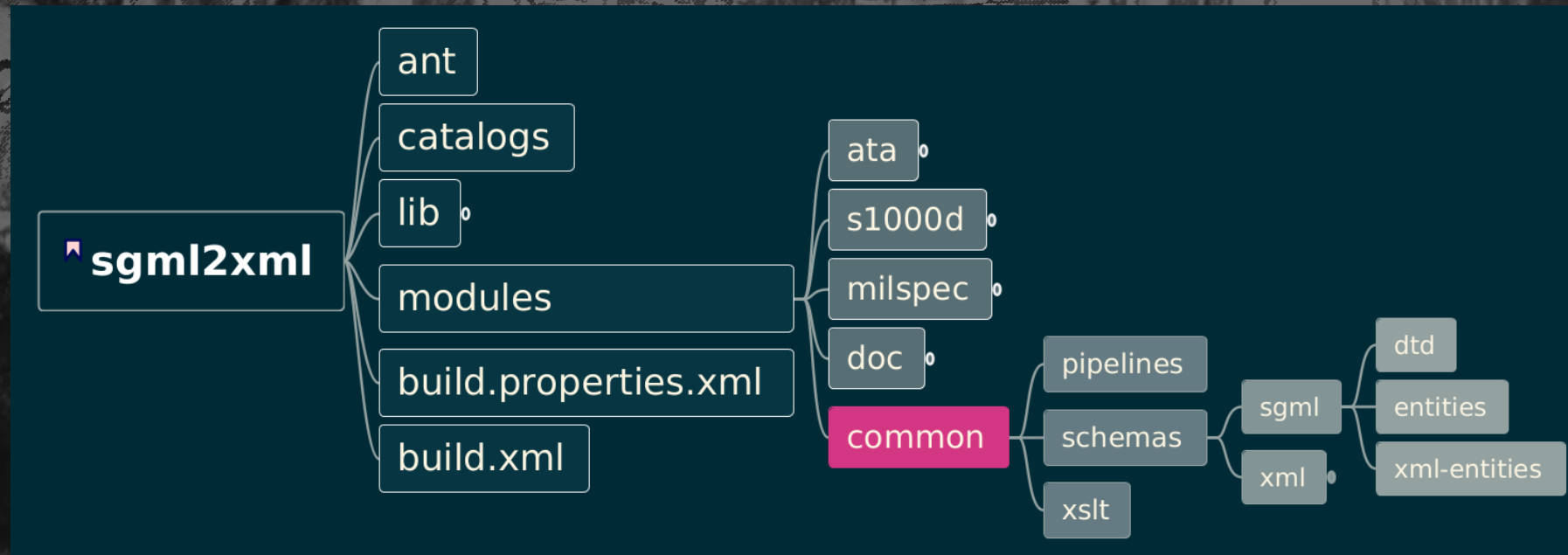


CREATIVE WORDS

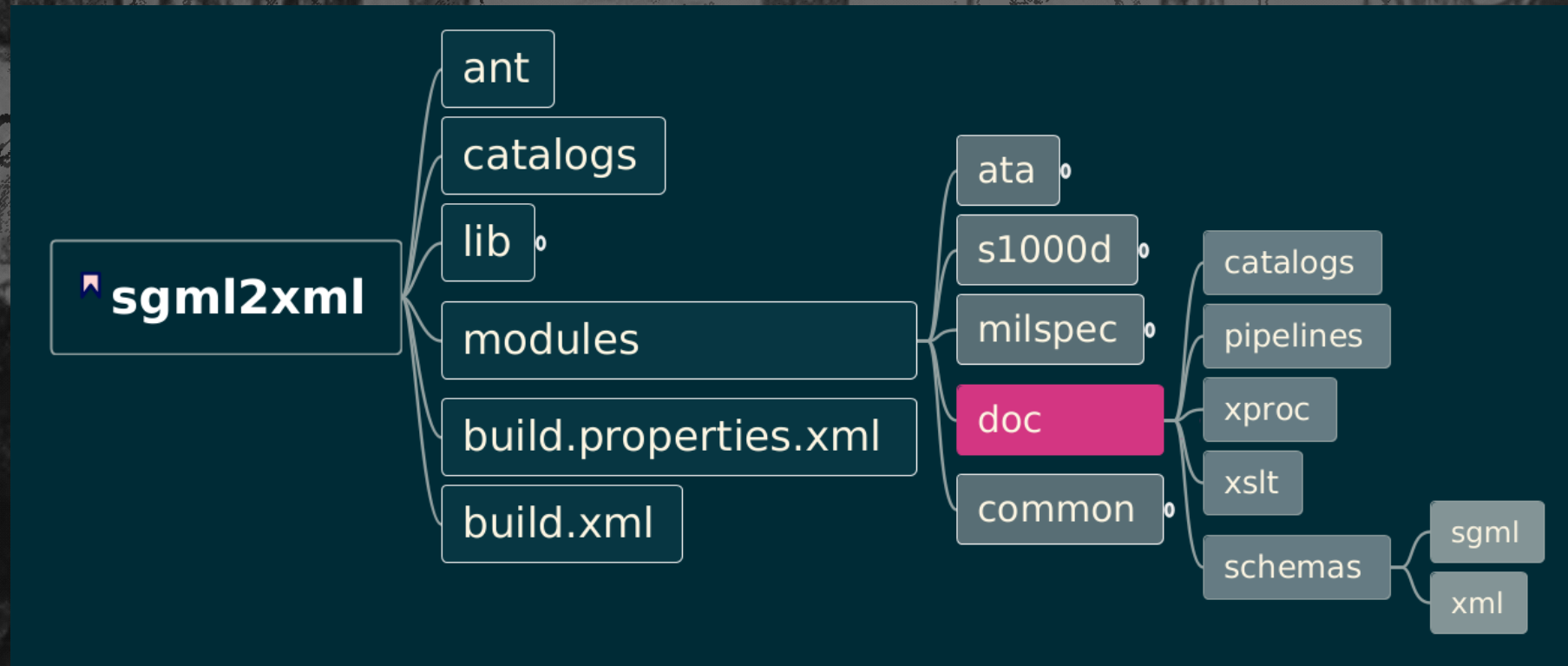
EACH DOCUMENT TYPE IS A MODULE

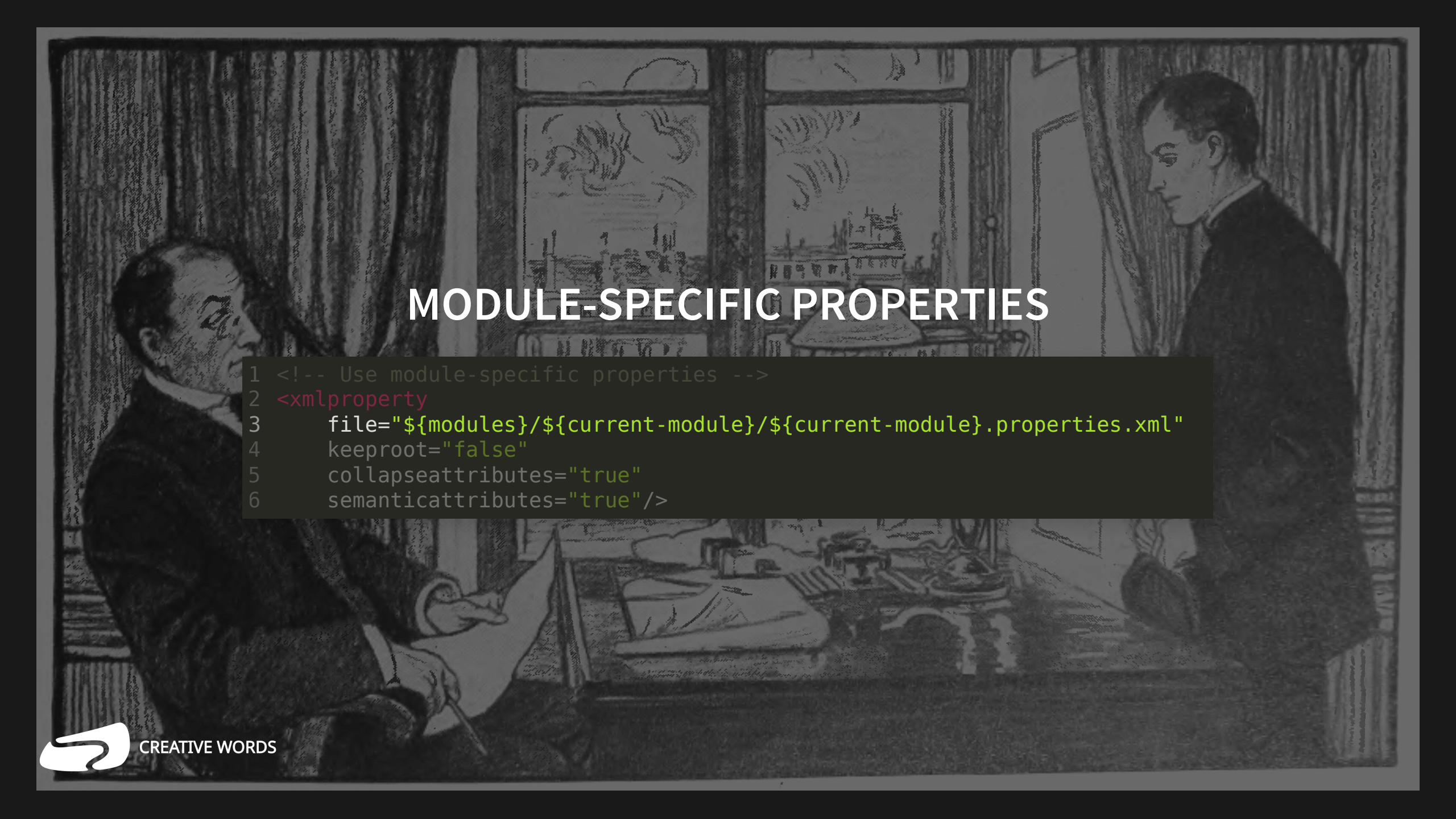


COMMON



doc MODULE





MODULE-SPECIFIC PROPERTIES

```
1 <!-- Use module-specific properties -->
2 <xmlproperty
3   file="${modules}/${current-module}/${current-module}.properties.xml"
4   keeproot="false"
5   collapseattributes="true"
6   semanticattributes="true"/>
```



ENTITY REFERENCES, PIPELINE MANIFESTS, SCHEMATRONS, DOCTYPE LOOKUP, ...

```
1 <properties>
2
3   <module location="${modules}/doc">
4     <schemas location="${module}/schemas">
5       <!-- SGML -->
6       <input location="${module.schemas}/sgml/catalog.txt"/>
7       <!-- SGML to XML conversion -->
8       <output location="${module.schemas}/sgml/catalog_sgml2xml.txt"/>
9       <!-- ATA resources -->
10      <resources>
11        <ent location="${module.schemas}/sgml/ISOent-declarations-xml.txt"/>
12      </resources>
13      <!-- SGML DOCTYPE lookup -->
14      <sgml-dotypes location="${module.schemas}/sgml/doctype-lookup.xml"/>
15    </schemas>
16
17    <!-- Schematron -->
18    <sch location="${module}/sch/test.sch"/>
19
20    <!-- XProc pipeline manifests and libs -->
21    <pipeline>
```



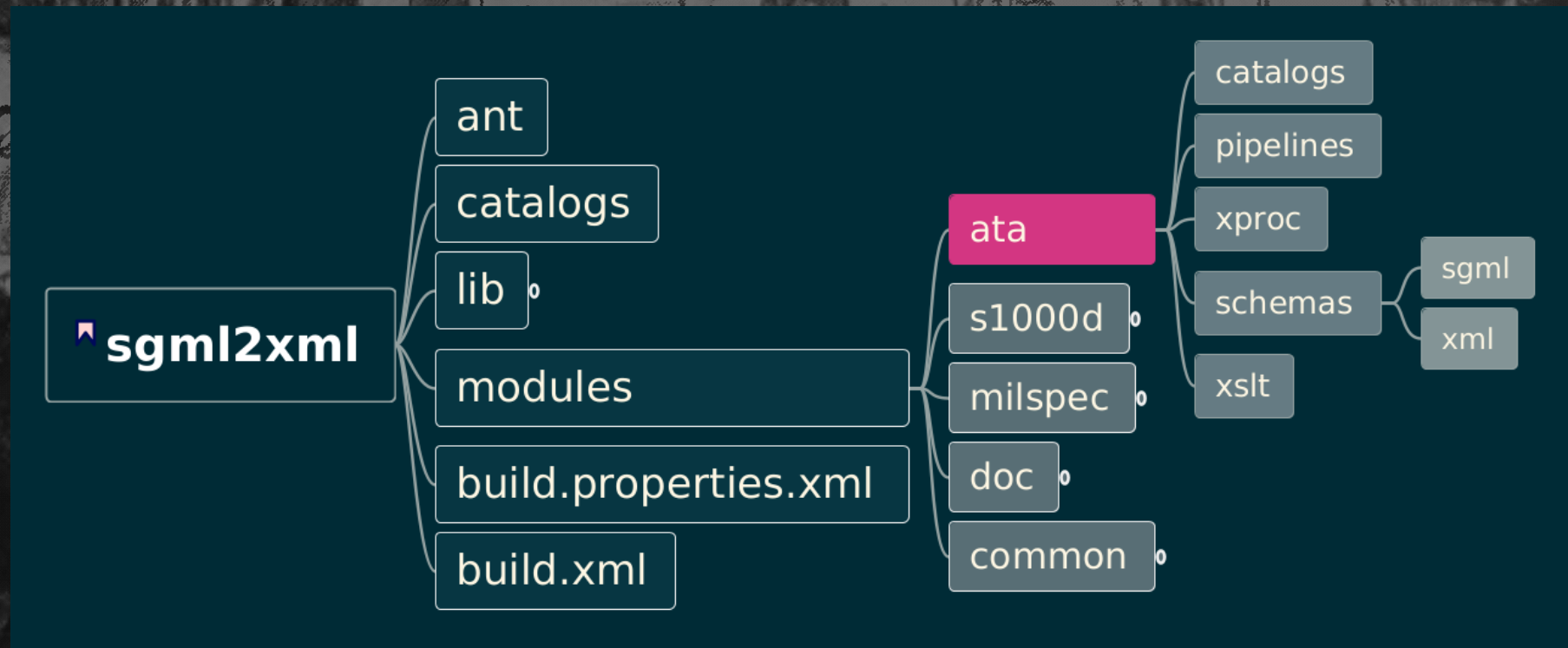


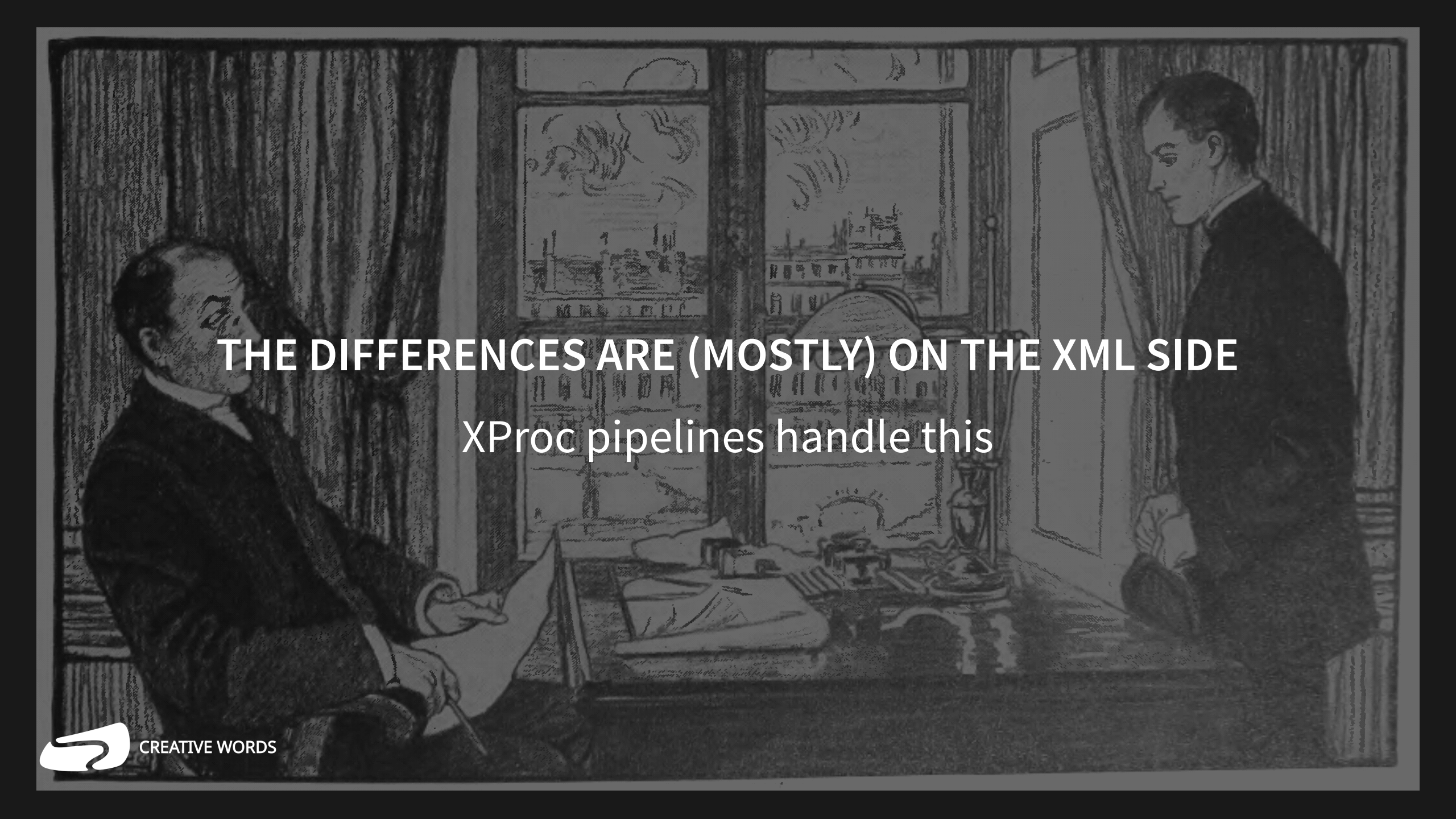
DOCTYPE RESOLUTION

```
<doctypes>
  <!-- doc example -->
  <doctype root="doc">
    <xslt-manifest>doc-xslt-manifest.xml</xslt-manifest>
    <xspec-manifest>doc-xspec-manifest.xml</xspec-manifest>
    <publicid>-//CWORD//DTD DOC TEST//EN</publicid>
    <systemid>doc.dtd</systemid>
  </doctype>
</doctypes>
```



ATA MODULE



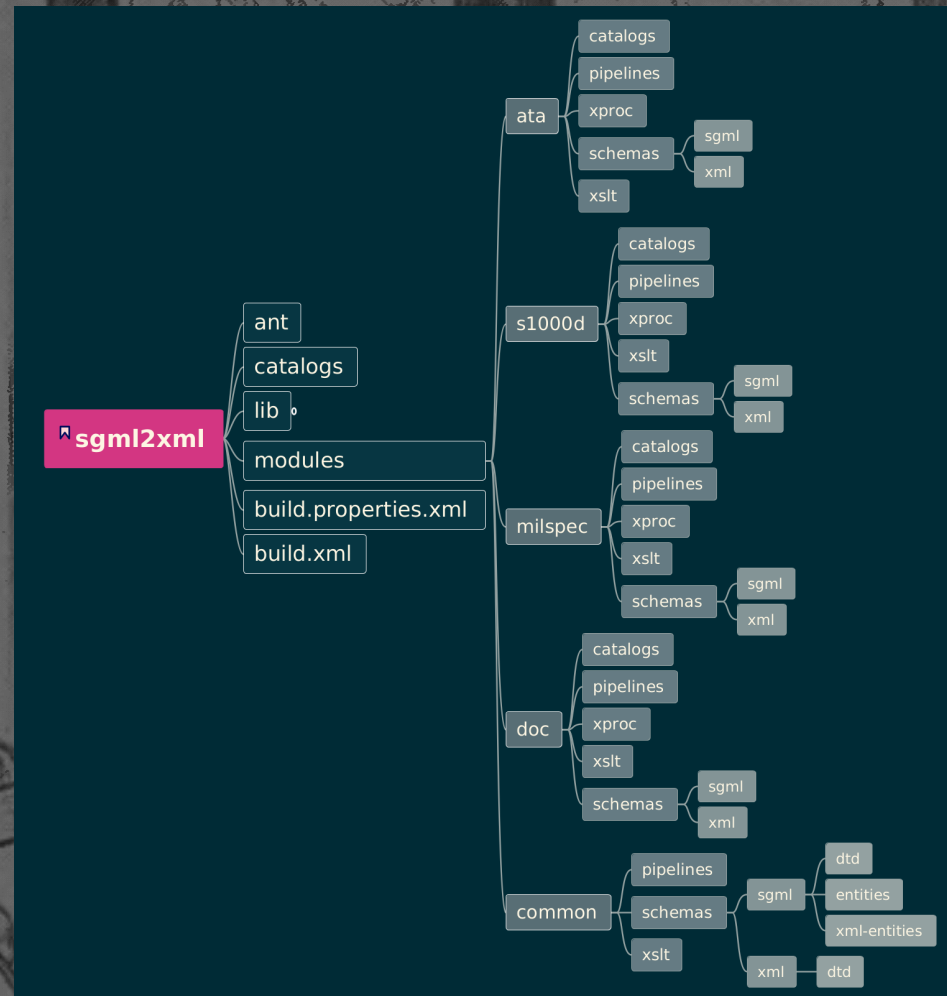


THE DIFFERENCES ARE (MOSTLY) ON THE XML SIDE
XProc pipelines handle this



CREATIVE WORDS

NOT PERFECT, BUT...



THERE AND BACK AGAIN (XML TO SGML)



CREATIVE WORDS

THE REVERSE PIPELINE

1. Strip XML DOCTYPE and generate SGML DOCTYPE with NDATA/NOTATION entries
2. Tweak models (e.g. replace @href references with entity-based graphic pointers)
3. Convert encoding (e.g. UTF-8 to ISO/ASCII)

THE CURIOUS CASE OF THE EMPTY ELEMENT

Expand EMPTY elements to explicit end tags before
spam normalisation, e.g

```
<graphic/>
```

to

```
<graphic></graphic>
```



IN CLOSING



CREATIVE WORDS

Édouard De Bièvre, The Compromise of Nobles, 1841.



WHY ANT?

- Inherited from original implementation
- **MorganaXProc-IIIse** (standard edition) lacks `p:run`
- Refactor towards pure XProc 3.x





THANKS TO

James Clark, Sheila Thomson, Nic Gibson,
Chris Pedoe, Henrik Mårtensson



CREATIVE WORDS

A detailed historical painting depicting a large group of men in 16th-century attire gathered in a grand hall with classical columns. The scene is filled with people, some standing and some seated at a long table in the foreground, engaged in various activities. The lighting is dramatic, with strong highlights and deep shadows.

<https://github.com/sgmlguru/sgml2xml-tng>



CREATIVE WORDS



QUESTIONS?



CREATIVE WORDS