Taller de marcado TEI

Nicolás Vaughan

Universidad de los Andes

n.vaughan@uniandes.edu.co

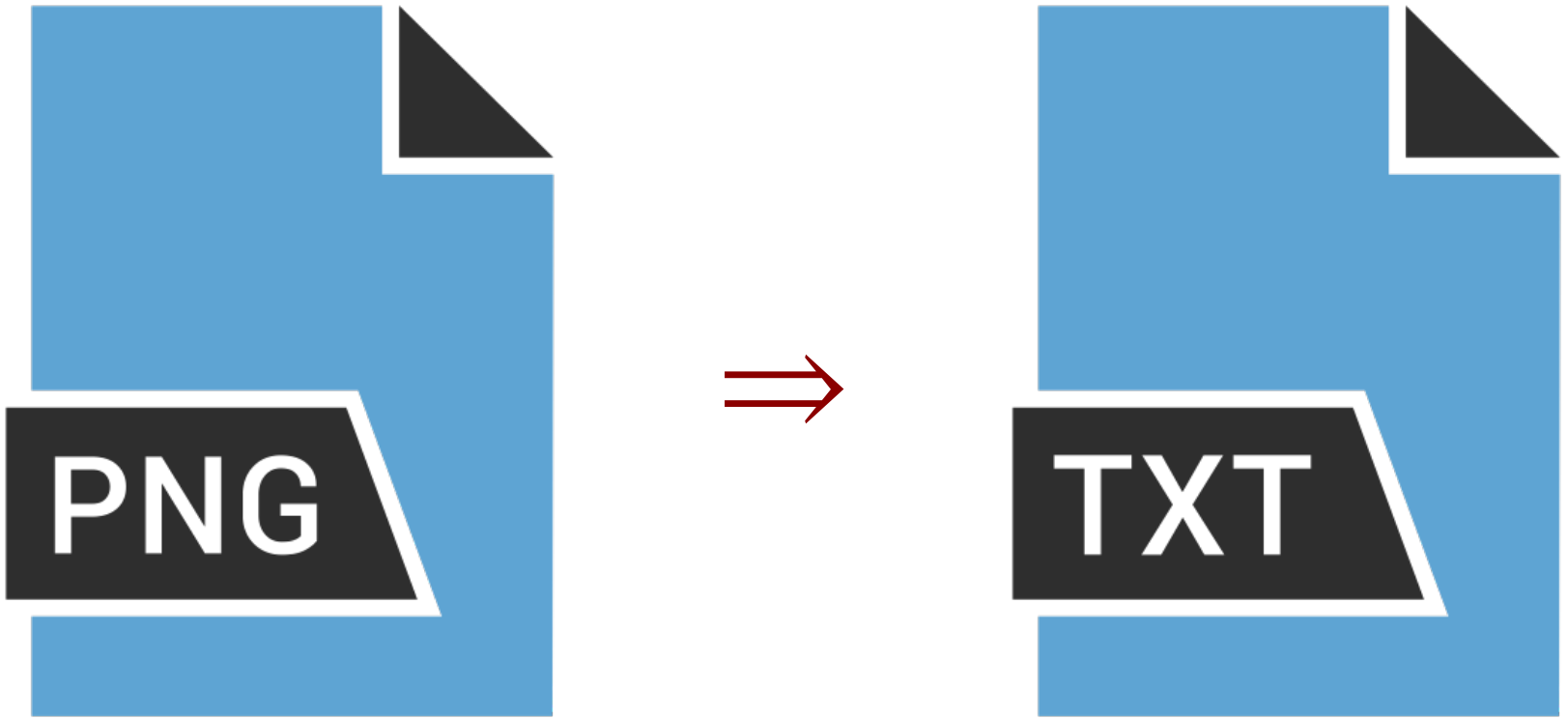
13 y 14 de abril de 2023

¿Qué significa 'digitalizar'?

1. *Escanear* — transformar una imagen analógica a un archivo digital gráfico (`.png`, `.tiff`, `.jpeg`, etc).

¿Qué significa 'digitalizar'?

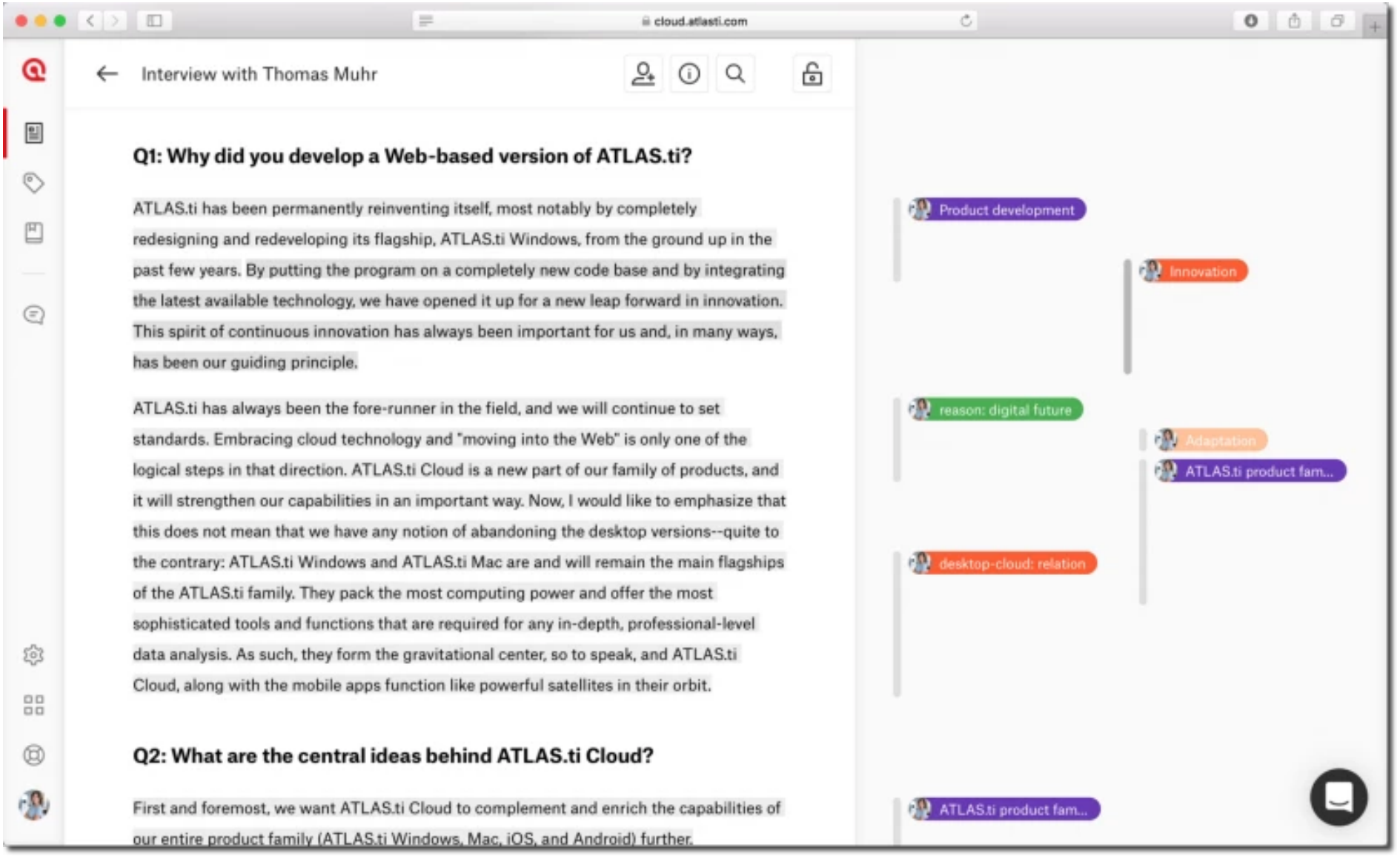
2. *OCR* — reconocimiento óptico de caracteres



¿Qué significa 'digitalizar'?

3. *Marcado / etiquetado (tagging)* — darle *significado* al texto

- representacional o gráfico (HTML, LaTeX, Troff, XSL-FO, etc.)
- semántico:
  - análisis y crítica textual
  - lingüístico
  - análisis cualitativo (qdap, ATLAS.ti, NVivo, etc.)
  - ...



Codificación categorías en ATLAS.ti

Marcado representacional en HTML

```
<!-- ... -->
<h1>Este es un título de nivel 1</h1>
<em>Este texto aparece en cursivas</em>
y <strong>este en negritas.</strong>

<br />

Una lista de viñetas:
<ul>
  <li>un ítem</li>
  <li>otro ítem</li>
  <li>otro ítem</li>
</ul>

<br />

<span style="color: red">
  Este texto va en rojo.
</span>
<!-- ... -->
```

Este es un título de nivel 1

Este texto aparece en cursivas y **este en negritas.**
Una lista de viñetas:

- un ítem
- otro ítem
- otro ítem

Este texto va en rojo.

Marcado semántico en TEI

```
<!-- ... -->
  <persName>Moctezuma Xocoyotzin</persName> nació en <date>1466</date>,
  y murió el <date>29 de junio en 1520</date>
  en <placeName>Tenochtitlan</placeName>, <placeName>México</placeName>.
<!-- ... -->
```

¿Qué es TEI?

TEI (*Text Encoding Initiative*) es una implementación del lenguaje de marcado XML diseñada para codificar o marcar semánticamente textos de diversas índoles.

Por su parte, XML (*Extensible Markup Language*) es un lenguaje de marcado general usado para codificar todo tipo de información.

Ejemplo de un documento XML

```xml
<?xml version="1.0"?>
<catalog>
  <book id="bk101">
    <author>Gambardella, Matthew</author>
    <title>XML Developer's Guide</title>
    <genre>Computer</genre>
    <price>44.95</price>
    <publish_date>2000-10-01</publish_date>
    <description>An in-depth look at creating applications
      with XML.</description>
  </book>
  <book id="bk102">
    <author>Ralls, Kim</author>
    <title>Midnight Rain</title>
    <genre>Fantasy</genre>
    <price>5.95</price>
    <publish_date>2000-12-16</publish_date>
    <description>A former architect battles corporate zombies,
      an evil sorceress, and her own childhood to become queen
      of the world.</description>
  </book>
</catalog>
```

XML: algunas definiciones

1. *Elementos* — son los pilares estructurales de todo documento XML. Se componen de:
   - una *etiqueta de apertura*
   - una *etiqueta de cierre*
   - un *contenido* (que puede ser otros elementos, texto o nada)
   - y opcionalmente unos *atributos* con sus *valores* correspondientes.

   Ejemplos:
   - `<persName>Moctezuma Xocoyotzin</persName>`
   - `<date when="2022-12-31">31 de diciembre de 2022</persName>`
   - `<date when="2022-12-31" calendar="juliano">diciembre 31</persName>`
   - ```
     <persName>
       <forename>William</forename>
       <surname>Shakespeare</surname>
     </persName>
     ```
   - `<lb></lb>` (o equivalentemente `<lb/>`)
   - `<lb n="3"></lb>` (o equivalentemente `<lb n="3"/>`)
     ```

XML: algunas definiciones

2. *Entidades*: XML contiene cinco caracteres que no pueden usarse literalmente sino solo por medio de una referencia:

`&quot`	"
`&amp`	&
`&apos`	’
`&lt`	<
`&gt`	>

3. *Padres, hijos, ancestros y descendientes*: si un elemento contiene otro elemento, el primero se denomina el *padre*, y el segundo el *hijo*. Un elemento puede tener muchos *ancestros* y muchos *descendientes*.
4. *Declaración*: está al principio de todo documento XML, identificándolo como tal: `<?xml version="1.0" encoding="UTF-8"?>`

XML: algunas definiciones

5. *Instrucciones de procesamiento*: van debajo de la declaración XML y especifican el modo como el documento debe ser validado semánticamente o procesado. Empiezan con <? y terminan con ?>.

   Por ejemplo, para validar un documento XML con el esquema de validación de TEI (más exactamente, el de TEI-all), debemos incluir lo siguiente:

```
<?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/
  relaxng/tei_all.rng" type="application/xml"
  schematypens="http://relaxng.org/ns/structure/1.0"?>
<?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/
  relaxng/tei_all.rng" type="application/xml"
  schematypens="http://purl.oclc.org/dsdl/schematron"?>
```

Reglas fundamentales de todo documento XML

1. Todo documento XML debe tener un único elemento raíz.

CORRECTO:

```
<biblio>
  <libro>
    <título>Cien años de soledad</título>
    <autor>Gabriel García Márquez</autor>
  </libro>
  <libro>
    <título>El coronel no tiene quien le escriba</título>
    <autor>Gabriel García Márquez</autor>
  </libro>
</biblio>
```

INCORRECTO:

```
<libro>
  <título>Cien años de soledad</título>
  <autor>Gabriel García Márquez</autor>
</libro>
<libro>
  <título>El coronel no tiene quien le escriba</título>
  <autor>Gabriel García Márquez</autor>
</libro>
```

Reglas fundamentales de todo documento XML

2. Todo elemento empieza con una etiqueta de apertura y cierra con una etiqueta de cierre.

CORRECTO:

```
<p>
  <title>Cien años de soledad</title>
  <author>Gabriel García Márquez</author>
</p>
```

INCORRECTO:

```
<p>
  <title>Cien años de soledad
  <author>Gabriel García Márquez</author>
</p>
```

Reglas fundamentales de todo documento XML

3. Todo elemento debe ser apropiadamente anidado.

CORRECTO:

```
<p>
  <q>Esta es una cita</q>
</p>
```

INCORRECTO:

```
<p>
  <q>Esta es una cita</p>
</q>
```

4. Los nombres de los elementos no pueden empezar con 'xml', números o puntuación (excepto '_').

CORRECTO:

```
<author>
<_author>
```

INCORRECTO:

```
<01_author>
<"author">
```

Reglas fundamentales de todo documento XML

5. Los espacios en blanco (caracteres de espacio, de tabulador y de salto de línea) *no* son significativos. XML suele tragarse los espacios múltiples.

   Esto:

```
<p>
  <title>
    Cien       años       de      soledad
  </title>
                    </p>
```

   es equivalente a esto:

```
<p><title>Cien años de soledad</title></p>
```

Validez sintáctica y semántica

- Un documento XML es *sintácticamente* válido si cumple con las reglas anteriores.
- Un documento XML es *semánticamente* válido si cumple con las reglas de un esquema de validación.
  - Un documento XML-TEI es semánticamente válido si cumple con las reglas prescritas por el consorcio TEI sobre el tipo de elementos (y sus atributos) y las relaciones existentes entre ellos.
  - Por ejemplo, que la raíz de todo documento debe ser el elemento `<TEI xmlns="http://www.tei-c.org/ns/1.0">`.
  - Y que dicho elemento debe tener obligatoriamente dos elementos hijos: `<teiHeader>` y `<body>`.
  - Y que el elemento `<p>` puede tener algunos atributos (e.g. `ana`, `cert`, `copyOf`, etc.), pero no puede tener otros (e.g. `type`).
  - Y así sucesivamente.

`https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html`

Software que usaremos

- El editor gratuito Visual Code Studio:
  `https://code.visualstudio.com/`
- Dos extensiones para ese editor:
  1. Scholarly XML:
     `https://marketplace.visualstudio.com/items?itemName=raffazizzi.sxml`
  2. tei-publisher-vscode:
     `https://marketplace.visualstudio.com/items?itemName=e-editiones.tei-publisher-vscode`