Gestión y análisis de datos, imágenes y texto

Introducción a TEI

Nicolás Vaughan

Universidad de los Andes

n.vaughan@uniandes.edu.co

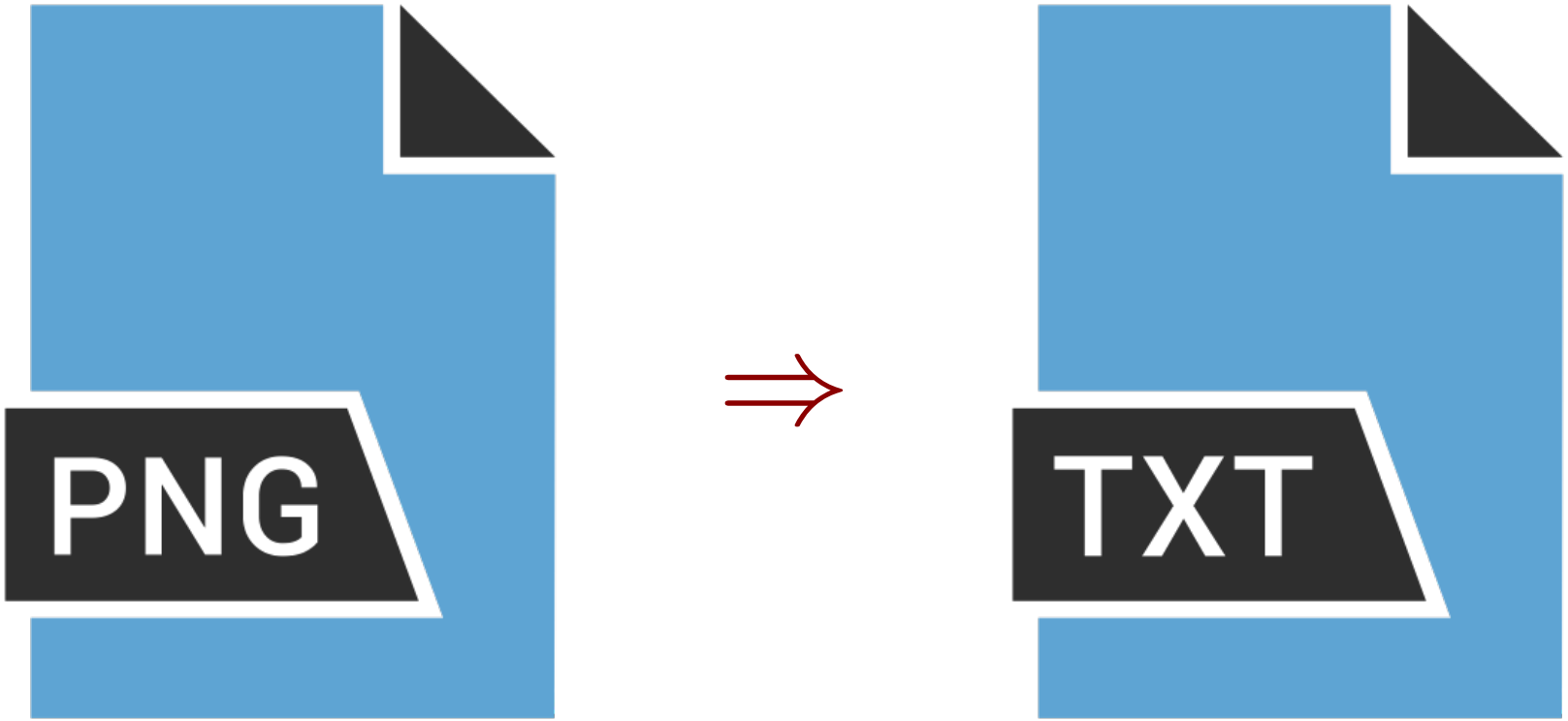
2025-08-27

¿Qué significa 'digitalizar'?

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

¿Qué significa 'digitalizar'?

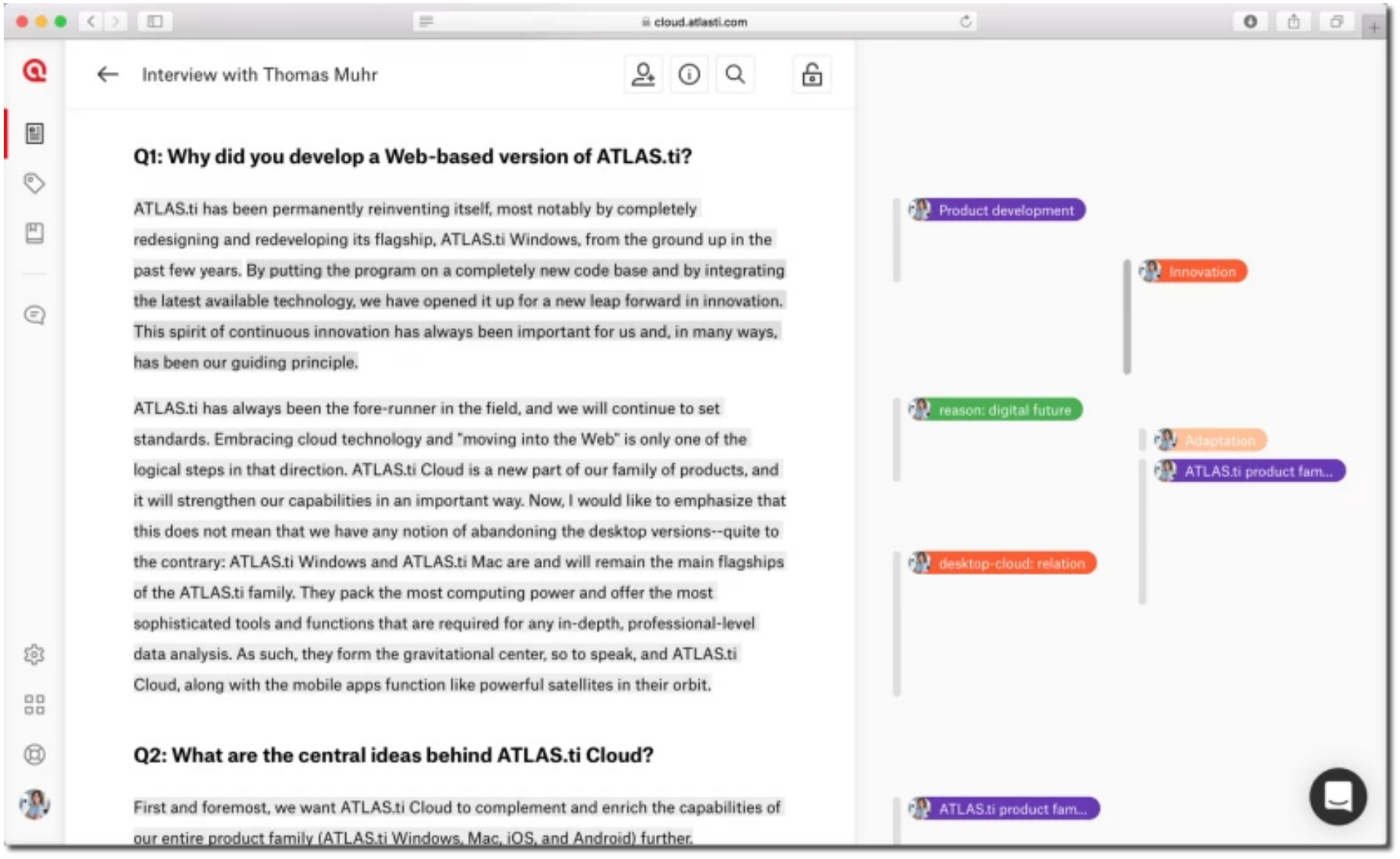
2. *OCR* — reconocimiento óptico de caracteres



¿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

El marcado: ¿para qué?

- Para poder *representar* correctamente su contenido (en una pantalla, en un papel, etc.)
- Para poder *interpretar* correctamente su contenido

En cualquier caso, es importante que el marcado sea *procesable por el computador* (*machine-readable*).

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 un 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;`	"
`&`	&
`&apos;`	'
`<`	<
`>`	>

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 un documento XML, identificándolo como tal: `<?xml version="1.0" encoding="UTF-8"?>`

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 un 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 un 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 un 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 un 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*.
  - Para nuestro caso, 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

El flujo de trabajo

1. La transcripción del documento (que puede ser manuscrito o impreso, en cuyo caso se puede usar OCR)
2. Codificación en TEI
3. Transformación del documento TEI resultante para su procesamiento, análisis, reutilización, etc.

   Por ejemplo:

   - TEI + XSLT → XML
   - TEI + XSLT → (X)HTML
   - TEI + XSLT → texto plano
   - TEI + XSLT → XSL-FO → PDF
   - TEI + XSLT → LaTeX → PDF
   - TEI + TeiPublisher → aplicación web
   - TEI + CETEIcean → aplicación web
   - TEI + XPath o XQuery → búsquedas estructuradas de información
   - …

Repositorio del taller

https://github.com/nivaca/taller-tei-2023

Bajemos la plantilla básica de TEI.

Algunos elementos comunes de TEI

- `<p>`: párrafos
- `<ab>`: bloques anónimos de texto
- `<q>`: texto entre comillas
- `<title>`: título de algún documento u obra
- `<name>`: elemento genérico de nombre de persona, institución, etc.
- `<persName>`: nombre de persona
- `<placeName>`: nombre de lugar
- `<date>`: fecha
- `<head>`: encabezado de una parte del texto
- `<list>`: lista
  - `<item>`: elemento en una lista
- `<cit>`: cita bibliográfica estructurada
  - `<quote>`: texto citado
  - `<bibl>`: entrada bibliográfica
- `<ref>`: referencia interna o externa

Algunos elementos comunes de TEI

Divisiones superiores del `<text>`

- `<frontmatter>`: contiene las páginas o información preliminar del documento (epígrafes, prólogos, introducciones, prefacios, etc.)
- `<mainmatter>`: contiende el texto principal
- `<backmatter>`: contiene las partes finales (apéndices, índices, etc.)
- `<div>`: división estructural genérica del documento (puede usarse el atributo `@type` para indicar si es de una parte, capítulo, sección, etc. y el atributo `@n` para indicar el número en su serie)

Hitos

- `<lb/>`: límite de línea
- `<pb/>`: límite de página
- `<cb/>`: límite de columna

Algunos elementos comunes de TEI

Correcciones e intervenciones editoriales

- `<add>`: texto añadido en el documento
- `<del>`: texto eliminado en el documento
  - `<subst>`: texto substituido en el documento (contiene un `<add>` y un `<del>`)
- `<sic>`: indica que el texto aparece tal cual en el documento, aunque el editor/codificador llama la atención sobre él
- `<corr>`: indica una corrección o intervención editorial
  - `<choice>`: puede contener una pareja `<sic>` y `<corr>` para indicar que van juntos
- `<abbr>`: indica una abreviatura en el documento
- `<expan>`: indica la expansión de una abreviatura en el documento
  - `<choice>`: puede contener una pareja `<abbr>` y `<expan>` para indicar que van juntos
- `<orig>`: indica que el texto aparece tal cual en el original
- `<reg>`: indica una normalización ortográfica
  - `<choice>`: puede contener una pareja `<orig>` y `<reg>` para indicar que van juntos

Correcciones e intervenciones editoriales

- `<unclear>`: indica que el texto es poco claro o ilegible (también se puede usar el atributo `@cert` para indicar el grado de certeza)
- `<gap>`: indica que hay una laguna en el texto (puede usar los atributos `@unit` para indicar la unidad de extensión (e.g. `caracteres`, `folios`) y `@extent` para indicar la cantidad)

  E.g. `<gap extent="2" unit="líneas"/>`

Correspondencia[I]

- `<stamp>`: contiene una descripción de un sello (`@type` puede especificar su tipo, e.g. `matasellos`, `estampilla`, etc.)
- `<opener>`: contiene la apertura de la comunicación
- `<dateline>`: contiene una descripción breve del lugar, tiempo, etc. de la producción de la comunicación
- `<address>`: es un elemento grupo que contiene varios elementos, como el genérico `<addrLine>` (que contiene una línea de dirección), u otros más específicos como `<street>` (la calle) o `<postCode>` (el código postal)
- `<closer>`: es un elemento grupo que contiene el cierre de comunicación (la despedida, la firma, etc.)
- `<signed>`: la firma (i.e. el nombre del remitente)

[I] https://tei-c.org/release/doc/tei-p5-doc/en/html/DS.html#DSOC

Otros elementos

- `<pc>`: puntuación
- `<g>`: caracteres y glifos
- `<note>`: indica que el texto marcado es una nota (`@type` puede indicar si es marginal, a pie de página, etc.; `@place` puede indicar la ubicación: al margen, arriba, abajo, etc.)
- `<seg>`: indica que el texto marcado es un segmento de otro elemento más grande.