EDICIÓN DIXITAL SEN ESTRÉS TECNOLÓXICO: DAS FONTES Á WEB

Helena Bermúdez Sabel

CONTIDOS

- Fluxos de traballo automatizados
- Introdución a XML e TEI
- Introdución a TEI Publisher
- Exercicio práctico

INFORMACIÓN PRELIMINAR

- Guía da sesión: **https://helenasabel.github.io/obradoiro.html**

FLUXOS DE TRABALLO AUTOMATIZADOS

TRANSCRICIÓN AUTOMÁTICA

- De imaxes a textos electrónicos (anotados)
- Ferramentas OCR e HTR. Algúns exemplos:
  - Extensións dos navegadores:
    - **Image Reader OCR**, **Copyfish**
  - Móbil:
    - **Google Keep**
  - PC:
    - **Tesseract**, **Zotero OCR**, **Transkribus**
    - Editores de PDF con tecnoloxía OCR: **Adobe Acrobat**, **Wondershare PDFelement Pro**
  - Plataformas web:
    - **Windows: Text Recognition**, **Soda PDF**, **Docsumo**, **Google Drive**, **Transkribus.ai**

TRANSCRICIÓN AUTOMÁTICA

- Dispoñibilidade de modelos pre-adestrados e algoritmos de adestramento fáciles de usar
- Conversión de arquivos en formatos de transcrición a arquivos semanticamente anotados (XML-TEI) a través de:
  - scripting (ex. **DiScholEd**, **LECTAUREP - Page2tei**, Transkribus)
  - aprendizaxe automática (*machine learning*): **Khemakhem et al. (2017)**, **Pagel et al. (2021)**

COLACIÓN AUTOMÁTICA

- Colación: comparación das lecturas de dous ou máis testemuños
- As ferramientas de colación automátican reciben como entrada a transcrición de cada testemuño; estas transcicións alíñanse entre si mediante un algoritmo de aliñación.

COLACIÓN AUTOMÁTICA

- Ferramentas:
  - **LERA - Locate, Explore, Retrace and Apprehend complex text variants**
  - **Variance viewer**
  - **CollateX**
  - **Hypercollate**
- Para saber máis:
  - **Obradoiro de introdución á colación automática (en inglés)**

ANOTACIÓN AUTOMÁTICA

- Ferramentas de procesamento da linguaxe. Tarefas máis habituais:
  - Lematización, etiquetaxe gramatical e descrición morfosintáctica
  - Recoñecemento e desambiguación de entidades nomeadas
  - Análise de sentimentos e emocións
  - Modelado de tópicos, extracción de palabras-chave
  - Anotación de elementos de versificación (rima, encabalgamento, escansión)
  - Anotación de figuras retóricas

ANOTACIÓN TEXTUAL

- **CATMA**
- **Semanticat**
- **TUSTEP**
- Anotación crítica
  - **Classical Text Editor**
  - **TEI Critical Apparatus Toolbox**

ENTORNOS VIRTUAIS DE INVESTIGACIÓN

- Sistemas en liña que facilitan a colaboración entre investigadoras/es. No contexto da edición, xeralmente inclúen ferramentas relacionadas co almacenamento de documentos e imaxes, anotación de imaxes e documentos, análise textual, visualización, etc.

Entornos de investigación virtuais, ferramentas para a publicación e explotación de corpus

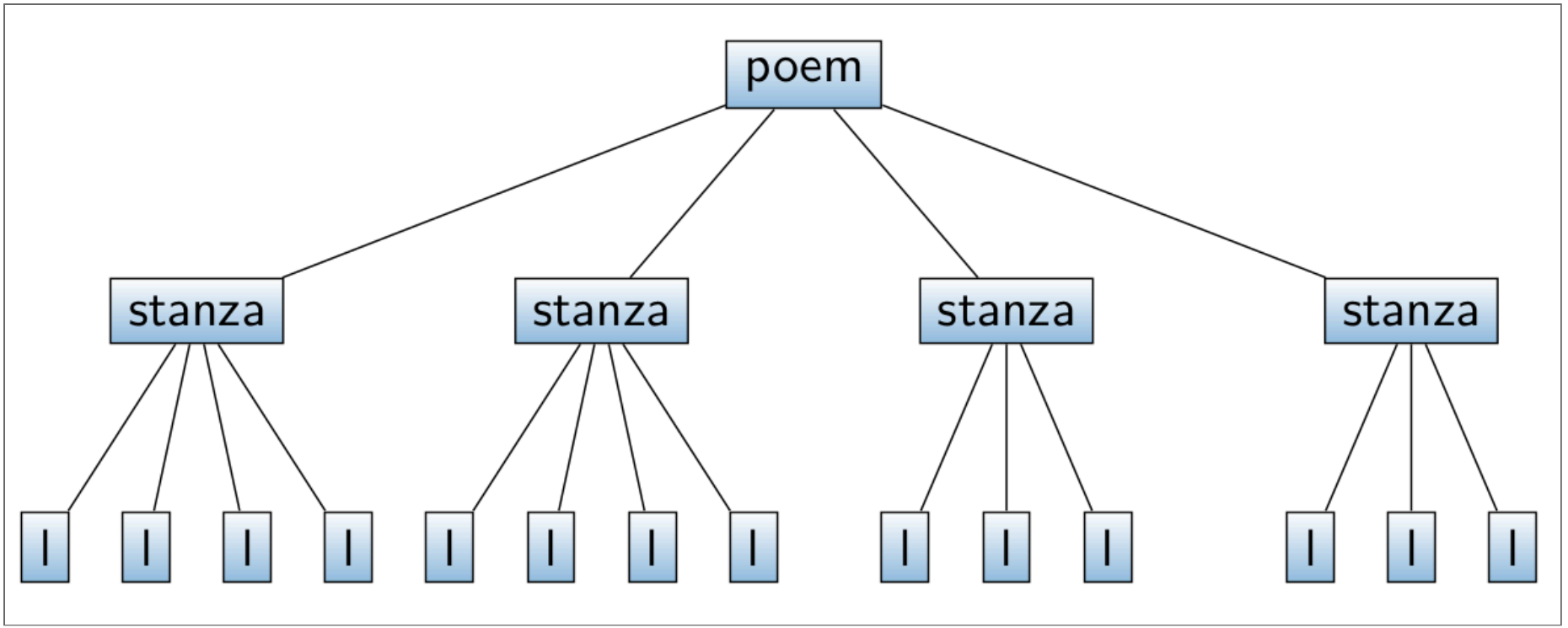
- **CETEIcean**
- **Ed. A Jekyll Theme for Minimal Editions**
- **Ediarium**
- **eLaborate**
- **FairCopy**
- **LEAF-Writer**
- **TEI Publisher**
- **TEITOK**
- **TEIViewer**
- **TextGrid**
- **Textual Communities**
- **TXM**

INTRODUCIÓN A XML E TEI

MARCAR IMPLICA…

… modelar a estrutura inherente dun texto e as súas propiedades semánticas a partir de:

- Xerarquías
- Estruturas ordenadas
- Linguaxe humana + linguaxe computacional


poem
stanza
stanza
stanza
stanza
l
l
l
l
l
l
l
l
l
l
l
l
l
l
l

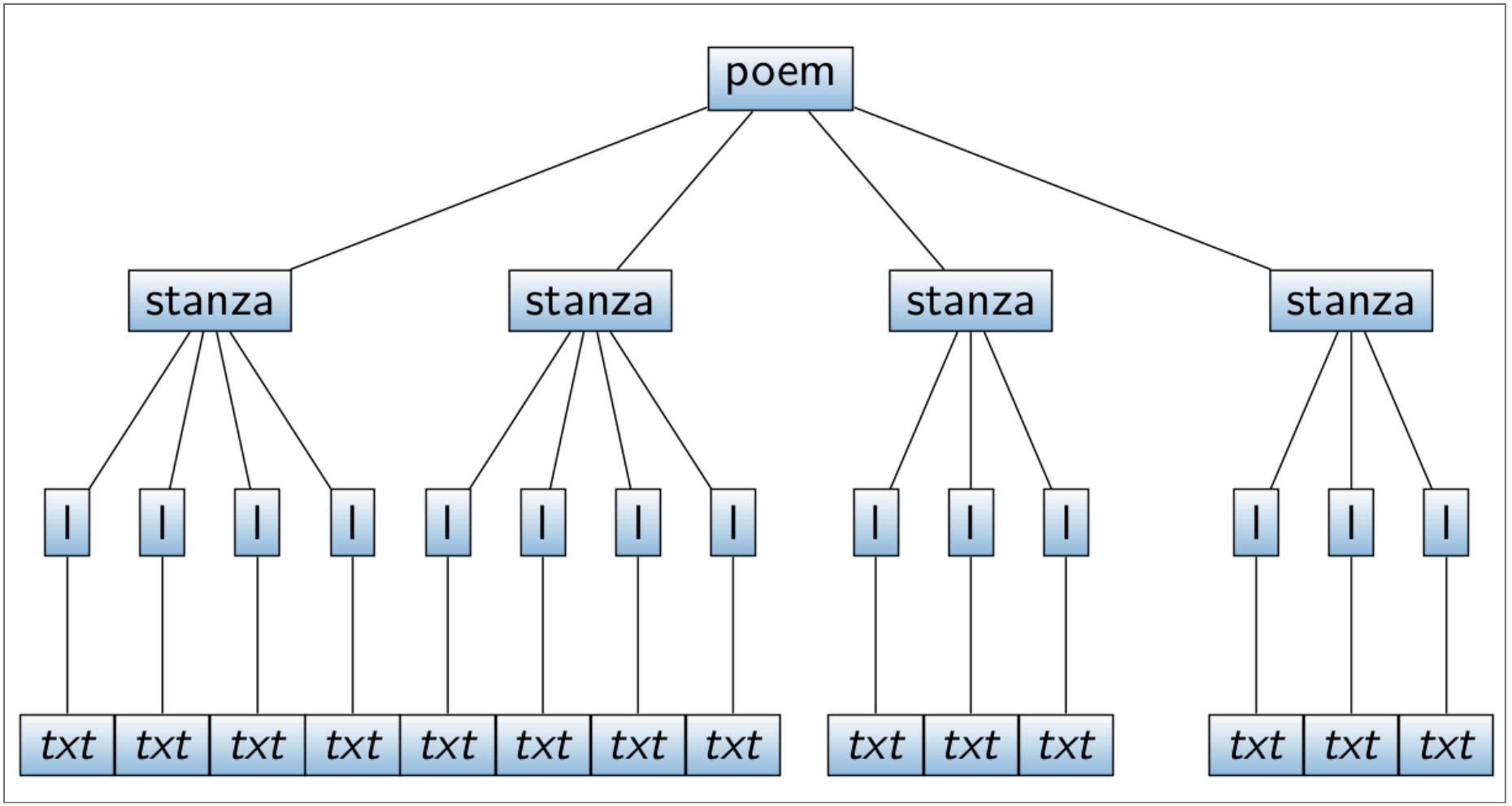
TIPOS DE MARCADO

- Descritivo
- Presentación
- Procedural
- **Multifacético**

QUE É XML?

- Formato para o almacenamento e transmisión de datos
- Definido polo World Wide Web Consortium (W3C)
- Uso moi extenso

```
<poem>
    <stanza>
        <l>Valenz Senher, rei dels Aragones</l>
        <l>a qi prez es honors tut iorn enansa,</l>
        <l>remembre vus, Senher, del Rei franzes</l>
        <l>qe vus venc a vezer e laiset Fransa</l>
    </stanza>
    <stanza>
        <l>Ab dos sos fillz es ab aqel d'Artes;</l>
        <l>hanc no fes colp d'espaza ni de lansa</l>
        <l>e mainz baros menet de lur paes:</l>
        <l>jorn de lur vida said n'auran menbransa.</l>
    </stanza>
    <stanza>
        <l>Vostre Senhier faccia a vus compagna</l>
        <l>per qe en ren no vus qal[la] duptar;</l>
        <l>tals quida hom qe perda qe gazaingna.</l>
    </stanza>
    <stanza>
        <l>Seigner es de la terra e de la mar,</l>
        <l>per qe lo Rei Engles e sel d'Espangna</l>
        <l>ne varran mais, si.ls vorres aiudar.</l>
    </stanza>
</poem>
```


poem
stanza
stanza
stanza
stanza
l
txt

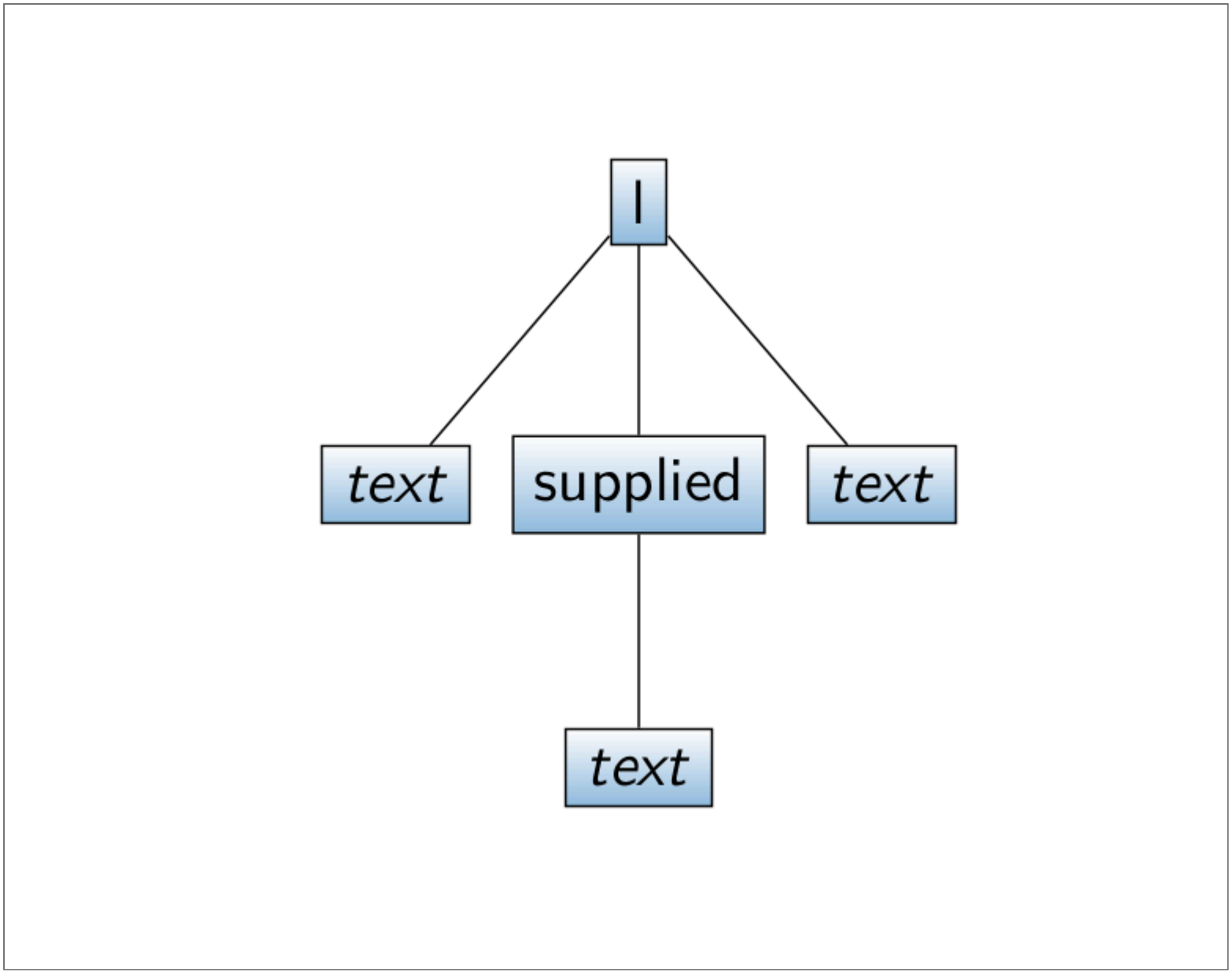
SINTAXE XML

- Elemento: `<l> </l>`
- Attributo: `<l met="3" rhyme="a">`
- Nós textuais:

```
<l met="3" rhyme="a">Frutales</l>
<l met="3" rhyme="b">cargados.</l>
<l met="3" rhyme="b">Dorados</l>
<l met="3" rhyme="a">trigales…</l>
```

```
<poem>
    <stanza type="quatrain">
        <l>Valenz Senher, rei dels Aragones</l>
        <l>a qi prez es honors tut iorn enansa,</l>
        <l>remembre vus, Senher, del Rei franzes</l>
        <l>qe vus venc a vezer e laiset Fransa</l>
    </stanza>
    <stanza type="quatrain">
        <l>Ab dos sos fillz es ab aqel d'Artes;</l>
        <l>hanc no fes colp d'espaza ni de lansa</l>
        <l>e mainz baros menet de lur paes:</l>
        <l>jorn de lur vida said n'auran menbransa.</l>
    </stanza>
    <stanza type="tercet">
        <l>Vostre Senhier faccia a vus compagna</l>
        <l>per qe en ren no vus qal[la] duptar;</l>
        <l>tals quida hom qe perda qe gazaingna.</l>
    </stanza>
    <stanza type="tercet">
        <l>Seigner es de la terra e de la mar,</l>
        <l>per qe lo Rei Engles e sel d'Espangna</l>
        <l>ne varran mais, si.ls vorres aiudar.</l>
    </stanza>
</poem>
```

```
<poem>
    <stanza type="quatrain">
        <l>Valenz Senher, rei dels Aragones</l>
        <l>a qi prez es honors tut iorn enansa,</l>
        <l>remembre vus, Senher, del Rei franzes</l>
        <l>qe vus venc a vezer e laiset Fransa</l>
    </stanza>
    <stanza type="quatrain">
        <l>Ab dos sos fillz es ab aqel d'Artes;</l>
        <l>hanc no fes colp d'espaza ni de lansa</l>
        <l>e mainz baros menet de lur paes:</l>
        <l>jorn de lur vida said n'auran menbransa.</l>
    </stanza>
    <stanza type="tercet">
        <l>Vostre Senhier faccia a vus compagna</l>
        <l>per qe en ren no vus qal<supplied>la</supplied>
            duptar;</l>
        <l>tals quida hom qe perda qe gazaingna.</l>
    </stanza>
    <stanza type="tercet">
        <l>Seigner es de la terra e de la mar,</l>
        <l>per qe lo Rei Engles e sel d'Espangna</l>
        <l>ne varran mais, si.ls vorres aiudar.</l>
    </stanza>
</poem>
```


l
text
supplied
text
text

SINTAXE XML

- A árbore XML ten unha única raíz, é dicir, un único elemento que contén todos os demais elementos
- Todos os contidos están delimitados
- Non pode conter os caracteres &, < (substituír polas entidades correspondentes `&` e `<`)
- Todos os elementos deben estar aniñados correctamente: sen solapamentos!

POR QUE XML?

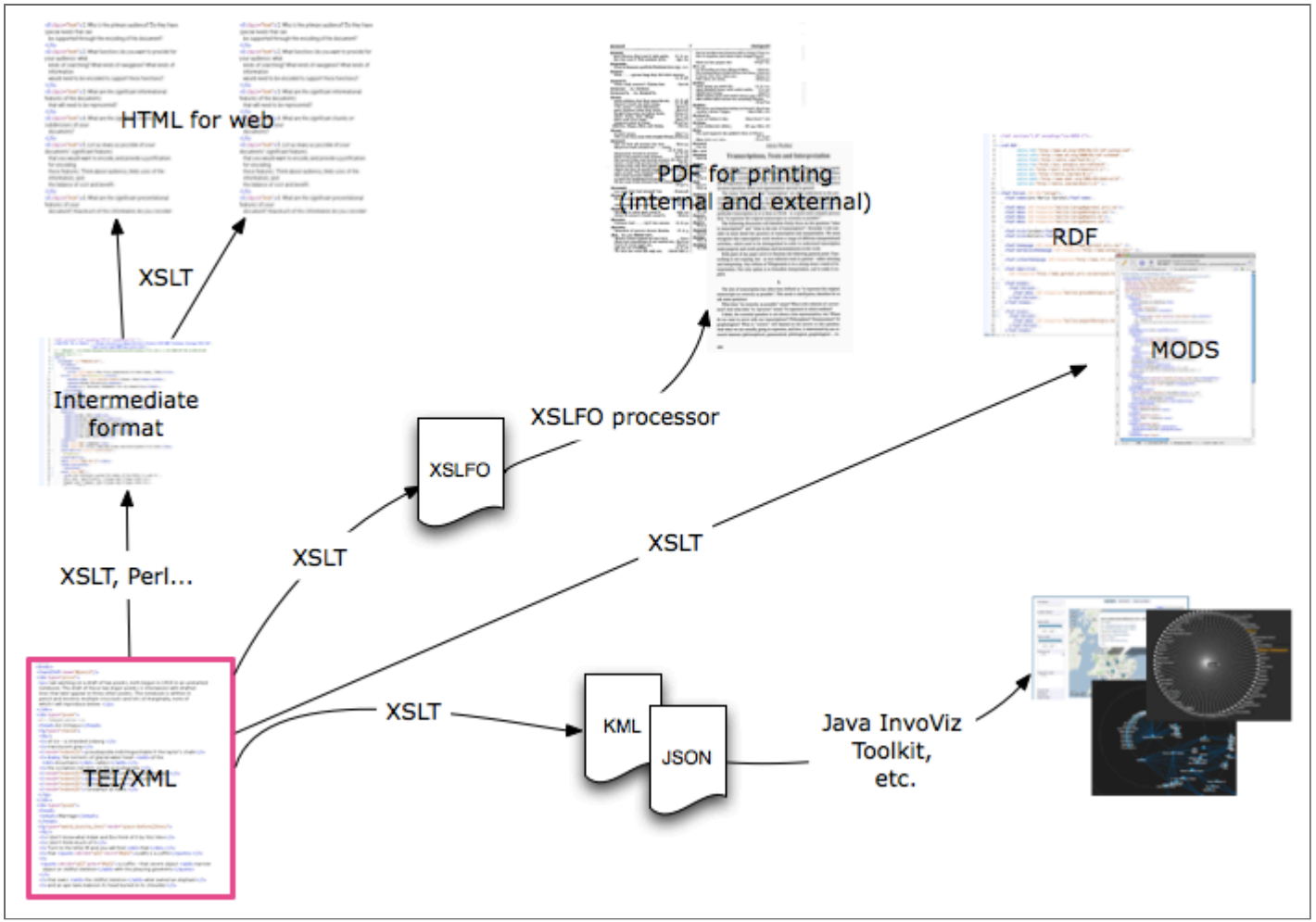
- Sintaxe fácil e simple
- Lexíbel
- Control da entrada e saída
- Software e hardware independentes
- Soportado por unha ampla gama de software (aberto + propietario)

XML E O MODELADO

- Modelado dun texto como **xerarquías ordenadas de obxectos de contido** (**Renear et al. 1993**)
  - Limitación: coexistencia de múltiples xerarquías lóxicas (o problema dos elementos superpostos)
  - Vantaxes do modelado:
    - As prácticas analíticas determinan frecuentemente xerarquías de obxectos (ou a descomposición en xerarquías adoita ser posible)
    - Lexíbel para persoas e máquinas
    - Sintaxe fácil

A FAMILIA DE ESTÁNDARES XML

- Linguaxes de esquema
- XPath
- XSLT
- XQuery
- SVG
- HTML
- KML
- XSL-FO, XForms, XProc, OOXML, OpenOffice.orgXML



Flanders (2018)

VÁLIDO VS. BEN FORMADO

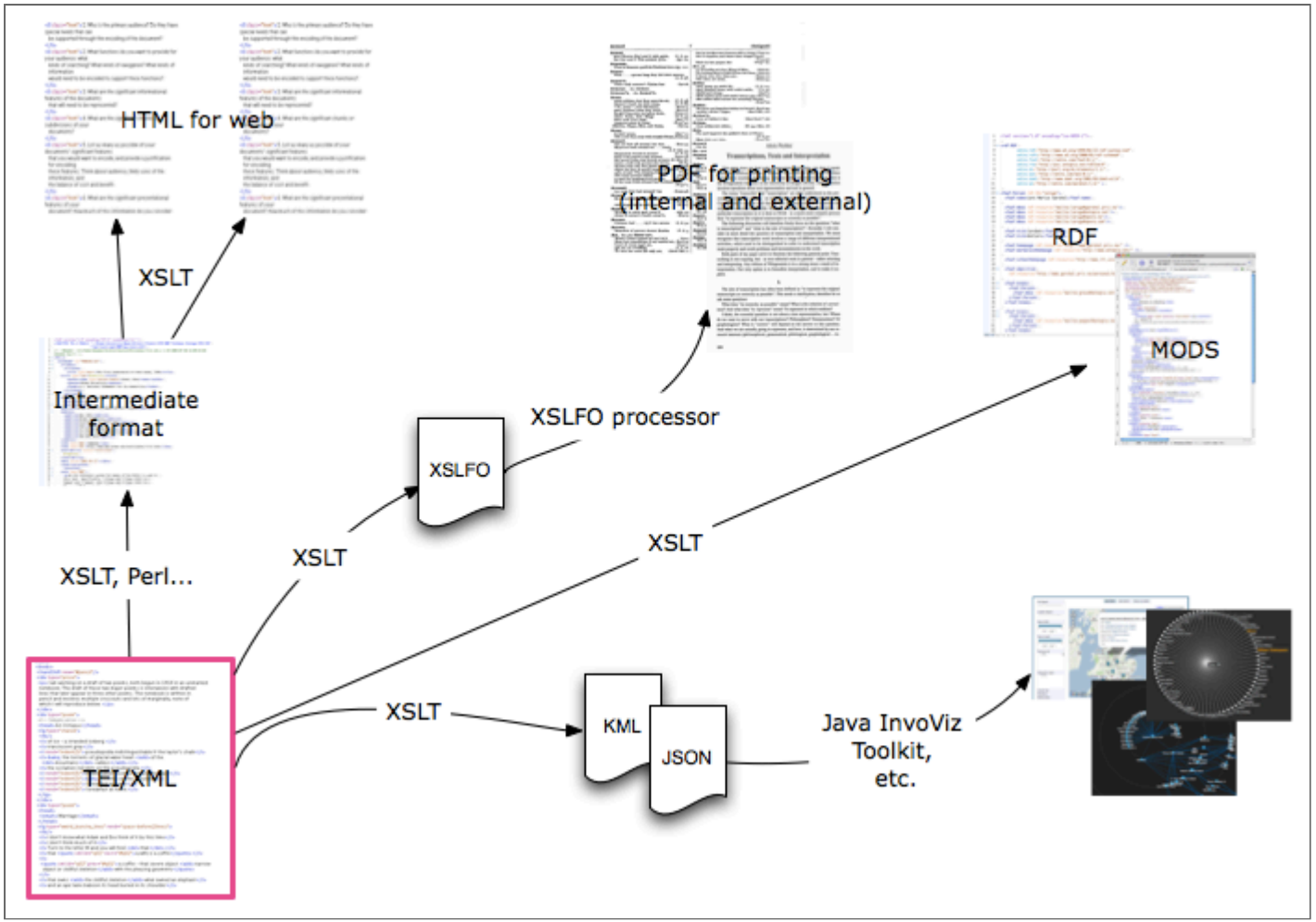
- Ben formado: cumpre coas regras de XML
- Válido:
  - Uso correcto do vocabulario: os elementos e atributos usados están dispoñíbeis nesa lingua
  - Uso correcto da gramática: os elementos utilízanse no lugar axeitado seguindo a orde definida

TEI COMO ESTÁNDAR XML

- Algunhas das vantaxes chave:
  - Interoperabilidade
  - Intercambio
  - Transferencia tecnolóxica
  - Eficiencia

QUE É A TEXT ENCODING INITIATIVE (TEI)

- Conxunto de pautas para codificar documentos culturais
- Consorcio Internacional que mantén e desenvolve as devanditas directrices
- Comunidade de proxectos e investigadores/as que implementan as directrices da TEI



Flanders (2018)

MODELADO EN TEI

- Definición do vocabulario
- Formalización das restricións
- Deseño do output (saída)

AS DIRECTRICES TEI

- Divididas en dúas partes
  - Capítulos (comunmente denominadas “prosa”)
  - Especificacións (*specs*)

VÁLIDO VS. BEN FORMADO

- Ben formado: cumpre coas regras de XML
- Válido:
  - Uso correcto do vocabulario: os elementos e atributos usados están dispoñíbeis nesa lingua
  - Uso correcto da gramática: os elementos utilízanse no lugar axeitado seguindo a orde definida
  - A validez garántese grazas á formalización dun esquema nalgunha das linguaxes de esquema

LINGUAXES DE ESQUEMA

- W3C XML Schema (XML Schema or XSD)
- Document Type Definition (DTD)
- REgular LAnguage for XML Next Generation (**RELAX NG**)
  - sintaxe XML
  - sintaxe compacta
- **Schematron**

- **TEI ODD**: un formato de especificación conforme á TEI-XML que permite customizar o vocabulario TEI dunha maneira que responde á **programación letrada**.

TEI ODD: ONE DOCUMENT DOES IT ALL

- Seleccionar módulos
- Eliminar elementos innecesarios
- Engadir novos elementos e/ou atributos
- Cambiar o nome dun elemento ou atributo
- Limitar os valores dun atributo
- Limitar a estrutura
- Manipular agrupacións funcionais de elementos
- Internacionalización (i-18n)
- Documentación
- Definir a saída a partir do **modelo de procesamento de TEI**

POR QUE TEI PUBLISHER?

TEI PUBLISHER

- Software libre de código aberto
- Basado en estándares abertos, con opcións por defecto que facilitan a anotación e publicación sen necesidade de programar
- Comunidade: **e-editiones** (organización de base co obxectivo de sustentar edicións a partir da promoción de estándares abertos e colaboración comunitaria)

DEMO

EXERCICIO PRÁCTICO

EDICIÓN DE *CONTOS DA MIÑA TERRA*