Grant Agreement Number **ECP 2006 DILI 510049**

ENRICH

ENRICH Corpus Analysis report

Deliverable number	***D6.1***
Dissemination level	*Public*
Delivery date	*26 September 2008*
Status	*Final*
Author(s)	*Elsa Zipstein*

e*Content*plus

This project is funded under the *e*Content*plus* programme,
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

Document Version Control

Version	Date	Change Made (and if appropriate reason for change)	Name and Institution of Commentator(s) or Author(s)
0.0	May 2008	The first draft	SYSTRAN S.A., France
0.1	July 2008	Second draft	SYSTRAN S.A., France
0.2	July 2008	New chapter "Information to be and not to be translated", metadata classification changed in chapter "Metadata content" from DC-based to TEI-based	AIP/TP
0.3	September	Final version	CCP

Document Review

Reviewer	Institution	Date and result of Review
Tomas Psohlavec	AIP	28 May 2008 Sent back with comments
Tomas Psohlavec	AIP	24 August 2008 (Accepted by Technical coordinator)
Zdeněk Uhlíř	NKP	31 August 2008 (Accepted for submission to EC)
Gabriella Lovasz	CCP	26 September 2008 (Doing changes required by the European Commission)

Document Signature/Approval:

Before the table of contents each document is to contain an approval signoff form.

Approved By (signature)	Date
	26 September 2008

Accepted by at European Commission (signature)	Date

Executive Summary

The Work package 6 aims to the integration of a tailored multilingual module via a user friendly sophisticated access.

Based on SYSTRAN's machine translation technology, this module will provide also terminology extraction and machine translation customization tools for the construction and retrieval of personalised metadata within the aim to create new multilingual digital documents and multilingual ontologies in Czech, Polish, Spanish, Portuguese, German, Italian, English, French, Danish, Hungarian, Russian and Serbo-Croatian.

One of the first objectives of that workpackage is to tailor the system so as to generate & provide personalisation resources for the translation of specific fields of the manuscripts description and/or documentation. The corpus analysis presented in that deliverable is an inventory of the Enrich manuscripts structure. The aim of that corpus analysis is to establish a structure that will allow us to create XSL Transformation stylesheets, usually used to transform a document described in an XML formalism into another XML formalism, to modify an XML document, or to publish content stored into an XML document to a publishing format (XSL-FO, (X)HTML…) for machine translation purposes. As the translation is driven by the document structure, it is easier to leverage this structure during the translation process and to keep it in the translated document, than with traditional document filters, which process the entire document linearly.

SYSTRAN Translation Stylesheets (STS) that will be the object of Deliverable 6.2 use XSLT to drive and control the machine translation of XML documents (native XML document formats or XML representations — such as XLIFF — of other kinds of document formats).

STS does not only provide a simple way to indicate which part of the document text is to be translated, but also enables the fine-tuning of translation, especially by using the structure of the document to help disambiguate natural language semantics and determine proper context. In that case, the STS would pass a title option to the translation engine. The stylesheet can activate Enrich-specific dictionaries for some parts of the document and can mark some expressions as not to be translated, in the same manner. The mechanism is implemented through XSLT extension functions. In particular, the stylesheet uses a `systran:translate` function to translate an XML fragment, and `systran: getValue/ systran:pushValue/ systran:popValue` functions for consulting and for setting linguistics options in the translation engine. Proper management of character properties is also provided so that, for instance, the translation of a phrase in bold font will appear in bold font, even if the phrase has moved within the translated sentence.
This process is highly customizable by the addition of new templates into the stylesheets.

Based on the current corpus structure, SYSTRAN has developed an xslt formalism allowing to the system to translate specific fields of the content metadata validated by the content partners.

CONTENT

I. Introduction

Content partners have provided us with content metadata sample so as to analyse the current content and develop the necessary xslt formalism allowing to translate specific content fields with respective translation parameters such as usage of customization resources, usage of specific linguistic and translation options.

This document aims to centralise the information provided about the corpus structure and analyze it for machine translation customization purposes and multilingual system tailoring.

Following the lines of the Description of Work, ENRICH, Annex 1, the main objectives of the Multilingual and user friendly sophisticated access are described in the Work package 6.

Work package Description					
Work package number :	**6**	**Start date:**	**3**	**End date:**	**24**
Work package title:	**Multilingual and user friendly sophisticated access**				

Objectives

This work package aims to the integration of a multilingual module via a user friendly sophisticated access: multilingual search application, multilingual forums, and multilingual ontology editor.
Based on SYSTRAN's machine translation technology, this module will provide also terminology extraction and machine translation customization tools for the construction and retrieval of personalised metadata within the aim to create new multilingual digital documents and multilingual ontologies in Czech, Polish, Spanish, Portuguese, German, Italian, English, French, Danish, Hungarian, Russian and Serbo-Croatian.

Description of work

Work package leader: SYS
Task 6.1 Multilingual access development (m0-m12)
Task leader: SYS
Task participants: NKP, AIP

The project will provide two types of multilingual access:

- via the API integration in the data retrieval interface associated or independent of a multilingual search.
- a dedicated translation interface where ENRICH expert users can fine-tune dynamically the machine translation tools thanks to adapted linguistic tools for terminology extraction and translation post-editing and customization. The parameters and resources constructed will be automatically taken into account by the API in the access presented above

Task 6.2 Translation Stylesheet design and use (m3-m24)
Task leader: SYS
Task participants: NKP, AIP, CCP, NFC, NLF, IMI, ULV, SAM, CSH, DSP, NLI, BNE, BUTE, ULW
Activities:

• analysis of heterogeneity of metadata regarding machine translation.
• implementation of STS exploiting metadata information.
• cross-language validation of STS, optimization of translation parameters.
As far as the Metadata translation module implementation is concerned SYSTRAN will provide a fully customized Translation Stylesheet.
SYSTRAN Translation Stylesheets (STS) use XSLT to drive and control the machine translation of XML documents (native XML document formats or XML representations — such as XLIFF — of other kinds of document formats). STS will provide a simple way to indicate which part of the document text is to be translated, and will enable the fine-tuning of translation, especially by using the structure of the document to help disambiguate natural language semantics and determine proper context. Thanks to STS machine translation is considered as part of the authoring and publishing process: source documents can be annotated with natural language mark-up produced by the author, a mark-up which will be processed by STS to improve the quality of translation, the gateway to the automatic publishing of a multilingual website from a monolingual (annotated) source. The mechanism is implemented through XSLT extension functions for consulting and for setting linguistics options in the translation engine. SYSTRAN will deliver this xslt file in order to fine-tune the system according to the ENRICH xml data elements.

Task 6.3 VICODI implementation (m6-m24)
Task leader: SYS
Task participants: NKP, AIP,

• definition and homogenization of initial ontology applicable for this project
• specification of user-friendly web-interface for visualization of multilingual ontology - special interface for modification
• implementation of the web-interface
Based on previous experience in the visualization and contextualization of digital content (IST project VICODI) SYSTRAN technology has been implemented for the construction of multilingual ontologies. The Research Center for Information Technologies (FZI) constructed multilingual ontologies available under GNU Free Documentation License (FDL) thanks to the EU-funded IST project Vicodi http://www.vicodi.org/). Enrich will implement and use VICODI ontologies for the contextualization of the digital content.

(Inter-) Dependencies, milestones[1] and expected result

Based on the ENRICH Corpus Analysis (Month 6) SYSTRAN will build ENRICH Translation Stylesheet. After a quality assessment procedure and based on the evaluation results (Month 20) SYSTRAN will proceed to the finalisation of ENRICH Translation Stylesheet (Month 24).
The WP depends mainly on the results of WP3, but is also interrelated with WP4.
The feedback is necessary from WP7.

Deliverables

D 6.1 ENRICH Corpus Analysis report (Month 6), responsible partner: SYS
D 6.2 Personalised Translation Interface delivery report (Month 12), responsible partner: SYS
D 6.3 ENRICH Translation Stylesheet delivery report (Month 24), responsible partner: SYS
D 6.4 Vicodi Ontologies implementation report (Month 24), responsible partner: SYS

II. Corpus Structure overview

1. Metadata content

Following groups of metadata are used in the Content Partners primary metadata sources:

- information related to the resource (**resInfo**)
    - resource identification set of information (**resIdentifier**)
    - information describing the resource (**resDesc**)
        - intellectual content related information (titles, content summaries, or more detailed conten: rubrics, incipits, colphons, imprints etc.) (**resContents**)
        - intellectual and other responsibilities related to resource (primary and secondary responsibilities) (**resResp**)
        - description related to physical aspects of the resource (condition, decoration etc.) (**resPhysDesc**)
        - history of the resource (origin, provenance, acquisition) (**resHistory**)
    - additional information related to the resource (related bibliography, information on existing record history, custodial history, and other internal resource related administrative etc.) (**resAdditional**)
    - keywords and other added classification criteria (**resKeywords**)
- metadata related to the digitised document (**digidocInfo**)
    - structural metadata set related to the processed digital document (structural maps, logical maps, data files lists etc) (**digidocStruct**)
    - preservation set of information related to the processed digital document (information on process of digitisation, calibration information, DTDs) (**digidocTechDesc**)
- administrative and management information related to the existing data and metadata (**adminInfo**)
    - identification set of information for the data and metadata (**adminIdentification**)
    - responsibilities (metadata and data creator, contributor, publisher etc.) (**adminResp**)
    - rights statement (**adminRights**)

The above **res***, **doc*** and **admin*** information sets are segmented into more detailed partial information structures according to the practice of particular partners and the particular format used. These general level information groups originated as an disjunction of types of information produced by the Content Partners (based on the D2.2) and are more or less used by all of them (See D2.2 for further details – the delaiverable is prepared paralelly with D6.1).

Metadata may be deployed in a number of ways:

Embedding the metadata in the Web page by the creator or their agent using META tags in the HTML coding of the page.
As a separate HTML or XML document linked to the resource it is described in a database linked to the resource. The records may either have been directly created within the database or extracted from another source, such as Web pages.
The simplest method is to add the metadata as part of creating the page. To support rapid retrieval, the metadata should be harvested on a regular basis by the site robot. This is currently by far the most popular method for deploying Dublin Core.
Creating metadata directly in a database and linking it to the resource, is growing in popularity as an independent activity to the creation of the resources themselves. A separate XML document linked to the record.

Terminology
Consistent use of language with metadata descriptions can aid in the consistent discovery of resources. The primary tool for ensuring consistent language usage is via controlled vocabulary, including the use of thesauri. A number of metadata elements would benefit from controlled values. That would allow to create and Enrich-specific multilingual dictionary validated by the experts for the elements that need to be translated.

3. Content Formalism examples

The MASTER+ DTD used by NKP for Digitized manuscripts, old printed books, historical maps, and other historical materials can be consulted at

http://digit.nkp.cz/MMSB/1.1/msnkaip.xsd

```
<?xml version="1.0" encoding="UTF-8" ?>
<!--  edited with XMLSPY v5 rel. 4 U (http://www.xmlspy.com) by Karel Kucera (private),
XML Spy 2006 sp2  -->
<!-- W3C Schema generated by XMLSPY v5 rel. 4 U (http://www.xmlspy.com)  -->
- <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  + <xs:element name="ACCESSORY">
  + <xs:element name="ADDR">
  + <xs:element name="ADDRESS">
    <xs:element name="ADDR_NAME" type="xs:string" />
  + <xs:element name="AGE">
    <xs:element name="ALTITUDE" type="xs:string" />
    <xs:element name="APERTURE" type="xs:string" />
  + <xs:element name="AUDIO">
    <xs:element name="AUTO_FOCUS" type="xs:string" />
    <xs:element name="BACK_LIGHT" type="xs:string" />
  + <xs:element name="BASIC_INFO">
  + <xs:element name="BASIC_PARAM">
  + <xs:element name="BIRTH_DATE">
  + <xs:element name="BITS_PER_COMP">
    <xs:element name="BRAND" type="xs:string" />
    <xs:element name="BRIGHTNESS" type="xs:string" />
  + <xs:element name="CAMERA_CAPTURE">
  + <xs:element name="CAMERA_INFO">
  + <xs:element name="CAMERA_LOCATION">
  + <xs:element name="CAMERA_SETTINGS">
  + <xs:element name="CAPTION">
```

Figure 1. Manuscriptorium metadata structure

The University of Oxford is using the TEI P5 formalism. The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. In addition to the Guidelines themselves, the Consortium provides a variety of supporting resources, including resources for learning TEI, information on projects using the TEI, TEI-related publications, and software developed for or adapted to the TEI.

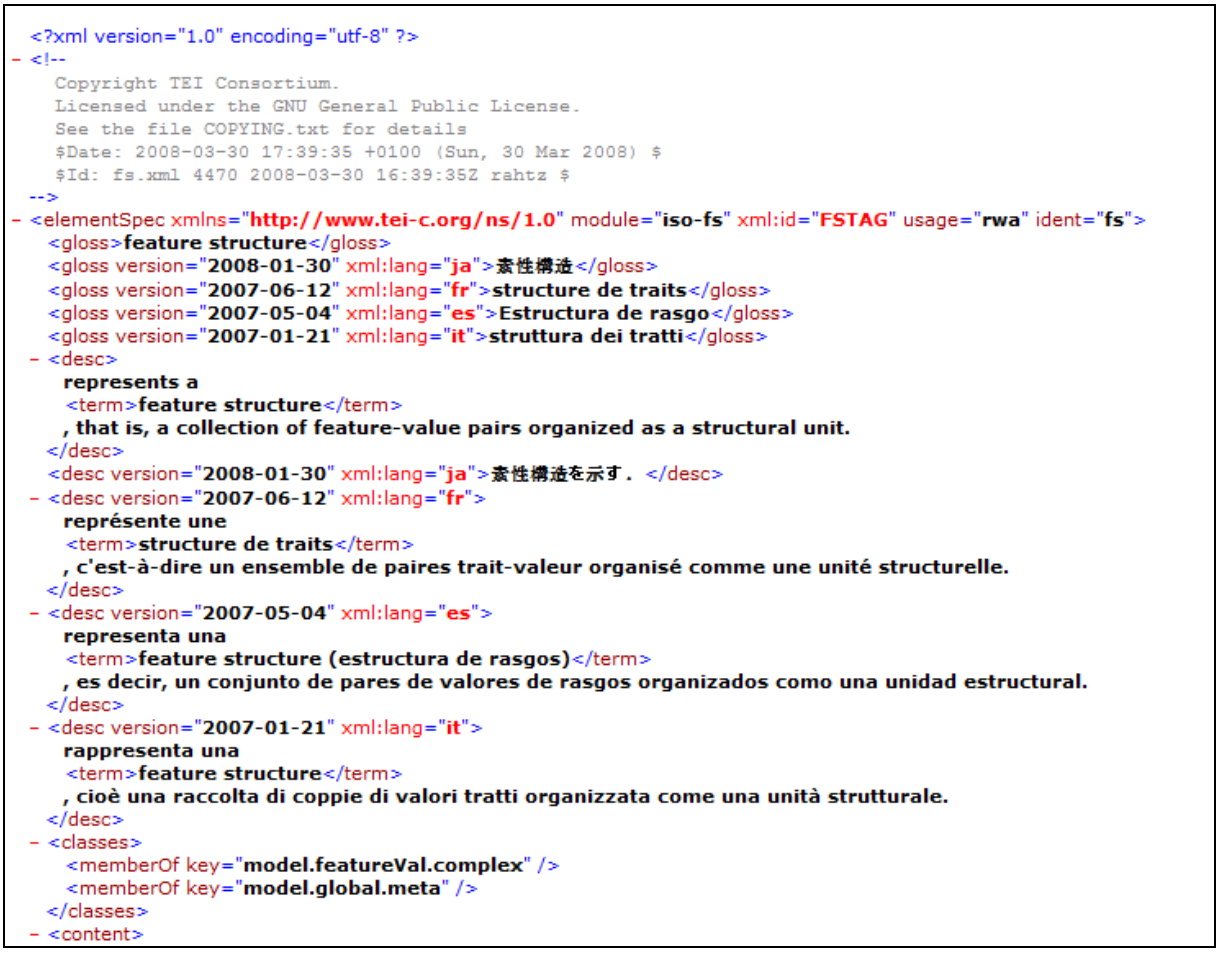


Figure 2. Oxford metadata structure

The description applied by the Servicio de Manuscritos e Incunables Biblioteca Nacional de España is following the formalism found below

```
- <record>
- <header>
  <identifier>oai:bibliotecadigitalhispanica.bne.es:169842</identifier>
  <datestamp>2008-03-17T11:42:18Z</datestamp>
  <setSpec>obrasm:obrff</setSpec>
  </header>
- <metadata>
- <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:title>De la clemencia</dc:title>
  <dc:creator>Séneca, Lucio Anneo</dc:creator>
  <dc:creator>Cartagena, Alonso de (1385?-1456)</dc:creator>
  <dc:type>manuscripttext</dc:type>
  <dc:type>Manuscrito</dc:type>
  <dc:publisher />
  <dc:date>S.XV</dc:date>
  <dc:language />
  <dc:description>El estoico Séneca escribe, a modo de proyecto de gobierno, esta
obra, supuestamente inacabada, entre el año 54 y el 55. Dirigida a Nerón, le ofrece
un modelo de lo que debe ser en el futuro, basado en el control del poder ilimitado
que posee mediante la clemencia, que no es tan sólo una disposición del alma, sino
una virtud activa cuyas consecuencias afectan a los seres humanos. La bondad
natural ayuda a evitar la corrupción del poder, pero este impulso natural debe
transformarse en un criterio de actuación. La clemencia es la única garantía de que
el soberano no se dejará arrastrar por las pasiones en el uso del poder, ya que
supone el ejercicio activo de su voluntad, así como la decisión ética de
autocontrolarse para conseguir el bienestar del pueblo, la credibilidad y la
seguridad. La obra está estructurada en dos libros, el segundo muy breve. Este
manuscrito es una traducción del siglo XV que puede atribuirse a Alonso de
Cartagena, quien sostiene que el libro II había sido elaborado en primer lugar, ya
que en él se encontraban las definiciones de clemencia</dc:description>
  <dc:description>El estoico Séneca escribe, a modo de proyecto de gobierno, esta
obra, supuestamente inacabada, entre el año 54 y el 55. Dirigida a Nerón, le ofrece
un modelo de lo que debe ser en el futuro, basado en el control del poder ilimitado
que posee mediante la clemencia, que no es tan sólo una disposición del alma, sino
una virtud activa cuyas consecuencias afectan a los seres humanos. La bondad
natural ayuda a evitar la corrupción del poder, pero este impulso natural debe
transformarse en un criterio de actuación. La clemencia es la única garantía de que
el soberano no se dejará arrastrar por las pasiones en el uso del poder, ya que
supone el ejercicio activo de su voluntad, así como la decisión ética de
autocontrolarse para conseguir el bienestar del pueblo, la credibilidad y la
seguridad. La obra está estructurada en dos libros, el segundo muy breve. Este
manuscrito es una traducción del siglo XV que puede atribuirse a Alonso de
Cartagena, quien sostiene que el libro II había sido elaborado en primer lugar, ya
que en él se encontraban las definiciones de clemencia</dc:description>
  <dc:description>Acorde de te escrevir, o Nero çesar, de la virtud que se llama
clemençia (h. 4v)... que lo tuerto e malo se enderesçe e torne derecho (h.
36v)</dc:description>
  <dc:description>Iniciales en azul y rojo con decoración de
rasgueo</dc:description>
  <dc:description>Calderones y títulos de capítulos en rojo</dc:description>
  <dc:description>Alvar y Lucía, Literatura medieval</dc:description>
  <dc:description>Grespi, Giuseppina, Traducciones castellanas de obras latinas e
italianas contenidas en manuscritos del siglo XV en las bibliotecas de Madrid y El
Escorial. Madrid, 2004</dc:description>
  <dc:description>Palabras sobre Séneca y su obra (h. 1). Carta del obispo de
Burgos, don Alonso llamado, al Rey don Juan, el segundo deste nombre (h. 2-3).
Introducción (h. 3-4v)</dc:description>

<dc:identifier>http://bibliotecadigitalhispanica.bne.es:1801/webclient/DeliveryMana
ger?pid=169842&custom_att_2=simple_viewer</dc:identifier>
  </oai_dc:dc>
  </metadata>
  </record>
```

Figure 3. Servicio de Manuscritos e Incunables Biblioteca Nacional de España metadata structure

According to the Biblioteca Nacional de España requirements the element to be translated at the first stage is the element <dc:description>.

The metadata of Diözese St. Pölten (DSP) are composed by 40.000 charters partially available via an MSQL database and little by little are converted to XML according to CEI schema.

The structural and descriptive metadata and part of the data (transcriptions) of the charters are organized in one model http://pcghw51.geschichte.uni-muenchen.de/UrkDTD/cei060122.xsd .

The hierarchy <chDesc> contains the metadata on structure and descriptive metadata.

Wroclaw University Library (WUL) will use the METS, the <mets:fileSec> and <mets:structMap> and will wrap WUL's internal XML using the <mets:mdWrap>.

Based on Wroclaw University Library (WUL) translation will be needed for the following elements

-main title

-title of the chapter/article

-keywords

-description of the document's content

-description of the content of the part (chapter)

-coverage: geographical place or region

-coverage: period, time

-details about the author, contributor

The Biblioteca Nazionale Centrale di Firenze (BNCF) digital collections metadata available for ENRICH project are in unimarcxml format.

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <rec>
  <lab>00884nez0 2200193 450</lab>
  <cf t="001">CF90010316</cf>
  <cf t="005">20061117183919.1</cf>
- <df t="100" i1="" i2="">
  <sf c="a">20061117f1780 |||||ita|01 ba</sf>
  </df>
- <df t="101" i1="0" i2="">
  <sf c="a">ita</sf>
  </df>
- <df t="102" i1="" i2="">
  <sf c="a">it</sf>
  </df>
- <df t="200" i1="1" i2="">
  <sf c="a">Pianta del Vicariato di Arcidosso</sf>
  </df>
- <df t="210" i1="" i2="">
  <sf c="c">Ultimo quarto del secolo XVIII</sf>
  </df>
- <df t="215" i1="" i2="">
  <sf c="a">carte: 1 altezza: 77,57 base: 60,32</sf>
  </df>
- <df t="300" i1="" i2="">
  <sf c="a">Descrizione esterna: Carta topografica con squadratura esterna.
China e acquarello su carta ruvida. Confini evidenziati con colori
all'acquarello, in porpora, giallo, verde; paludi, fiumi e, in generale,
l'idrografia, sono acquarellati in arrurro; orografia a monticelli in tinte
di grigio; toponomastica a penna.</sf>
  </df>
- <df t="610" i1="0" i2="">
  <sf c="a">Arcidosso</sf>
  </df>
- <df t="702" i1="" i2="1">
  <sf c="a">Giachi, famiglia (Antonio, Francesco, Luigi)</sf>
  <sf c="3">CF9V013688</sf>
  </df>
- <df t="801" i1="" i2="0">
  <sf c="a">IT</sf>
  <sf c="b">BNCF</sf>
  <sf c="c">20061117</sf>
  </df>
- <df t="950" i1="" i2="0">
  <sf c="a">Bibl. Nazionale Centrale di Firenze</sf>
  <sf c="d">CFA.I.13.103</sf>
  <sf c="f">A.I.13.103</sf>
  </df>
- <df t="956" i1="" i2="1">
  <sf c="a">BNCF0003496853</sf>
  </df>
  </rec>
```

Figure 4. Bibl. Nazionale Centrale di Firenze metadata structure

The following elements have to be translated:

df t = "3xx"

df t = "6xx"

where 3xx indicates all the tags beginning with 3 and 6xx indicates all the tags beginning with 6.

The University of Köln is using TEI with the following structure:

```
<?xml version="1.0" encoding="UTF-8"?>
<!ENTITY % a.global "id ID #IMPLIED
            n CDATA #IMPLIED
            authority CDATA #IMPLIED
            auth-range CDATA #IMPLIED
            lang (la | en | de | fr | it) #IMPLIED
            rend CDATA #IMPLIED">
<!ENTITY % e.global "bibl | date | note | ref | term | title">
<!ENTITY % e.layout "lb | pb | sub | sup | img | displayScript">
<!ENTITY % e.measures "height | width | depth | format">
<!ENTITY % e.textIdentifier "incipit | explicit | colophon | initium |
rubric">
<!ELEMENT CEEC (CEC+)>
<!ELEMENT CEC (TEIHeader, body?, addenda?, cover?)>
<!-- das Wurzelelement der xml-Dateien ist CEC; Veraenderungen an den
Daten werden der Einheitlichkeit
      willen allerdings an einer "Gesamt"-Datei vorgenommen, fuer die
das Wurzelelment CEEC
      vorgesehen ist -->
<!ELEMENT TEIHeader (fileDesc, profileDesc?, revisionDesc?)>
<!ATTLIST TEIHeader
      CodexId CDATA #REQUIRED
      lang (de | la | en | fr) #REQUIRED
>
<!ELEMENT abbr (#PCDATA)>
<!ATTLIST abbr
      type (withLine) #REQUIRED
      expan CDATA #IMPLIED
>
<!-- im Head eines CEEC-XML-Files wird nur 'abbr' erlaubt mit dem
Attribut 'expan', nicht aber
      umgekehrt auch 'expan' mit dem Attribut 'abbr' -->
<!ELEMENT accMat (#PCDATA | locus | ref)*>
<!ELEMENT acquisition (#PCDATA | person)*>
<!ATTLIST acquisition
      notBefore (1540 | 1861 | 1872) #REQUIRED
      notAfter (1540 | 1872) #IMPLIED
>
```

Figure 5. University of Köln metadata structure

4. Information to be and not to be translated

While providing the formalism samples the partners were also asked to state what part of the information they consider to be important for translation.

There were no partucilar guidelines or questionnaires prepared due the following:
due differencies between primary metadata structures (no universal metadata scheme is applicable until related parallel T3.1 and WP5 tasks are completed)
we wanted to avoid superabundant limitations by placing some preliminary structure limitations
Therefore there was a necessary generalisation of some of the answers in order to enable final common interpretation of the gathered information.

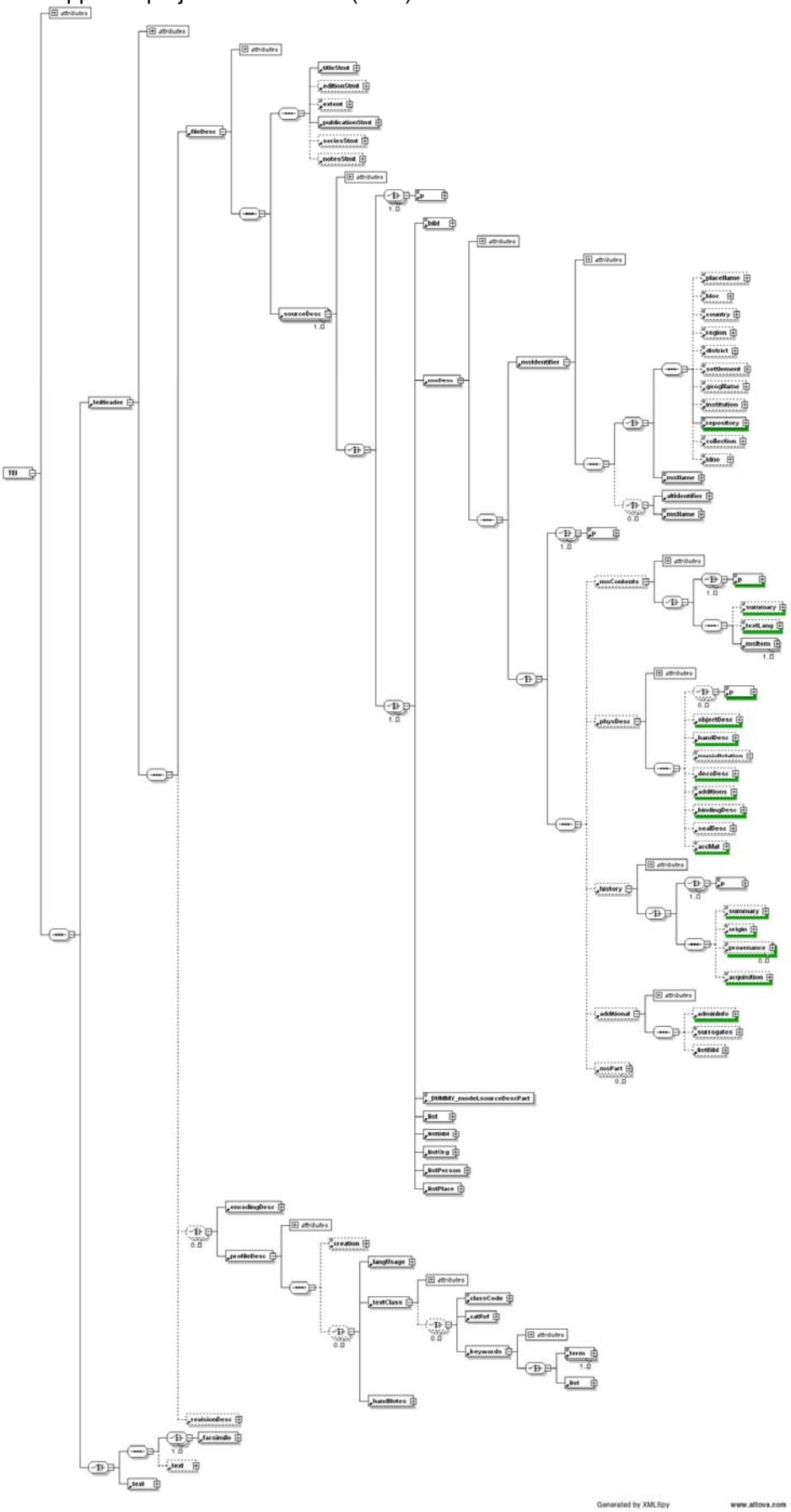
As noted above the final mapping between the numerous primary sources and the TEI P5 structure is not finished during preparation of this delivarable. In order to predict the final set of TEI P5 elements included into the tranlations we may predict a general supposed set of TEI P5 elements that will be used within Manuscriptorium. informations to be and not to be translated using the general sets of informations as described in the chapter „1. Metadata content“ and to project these information sets into the supposed TEI P5 structure.

NKP

NKP expressed the interest to translate the msDescription element, in particular:

msIdentifier/repository
msHeading/origDate
msHeading/origPlace
msHeading/textLang
msHeading/note
msContents/overview
msContents/note
msItem/summary
msItem/note
//decoNote
physDesc/form
physDesc/support
physDesc/layout
physDesc/msWriting
physDesc/decoration
physDesc/bindingDesc
physDesc/foliation
physDesc/additions
physDesc/condition
history/origin
history/provenance
history/acquisition
additional/availability
additional/custodialHist
additional/accMat

The supposed projection to TEI P5 (NKP):[1]



[1] See bigger version in Annex 1

OUCS

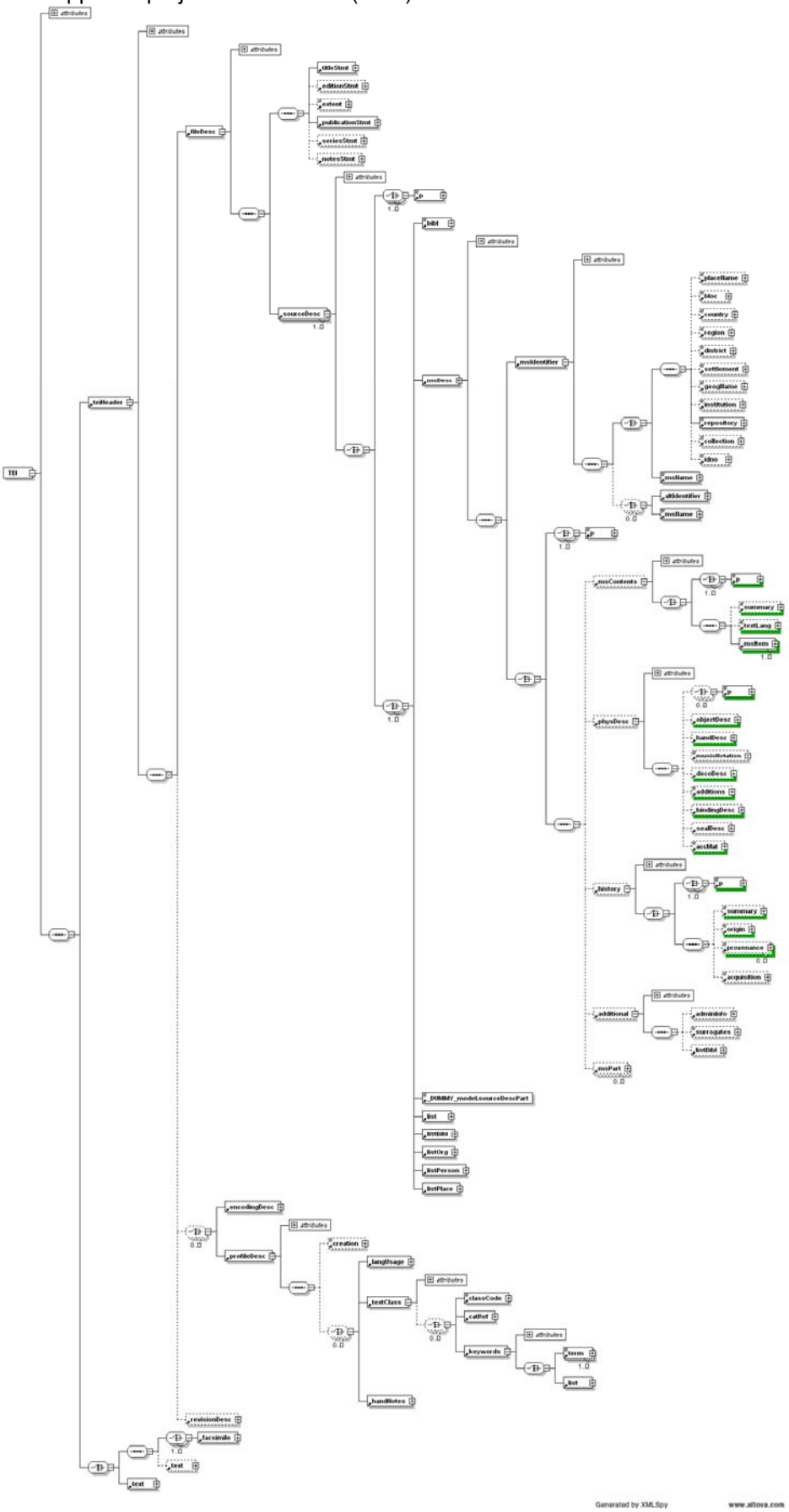
Oxford expressed the interest to translate only the contents of the <gloss> and <desc> elements, but not if they have an xml:lang attribute and not to translate the content of <term> elements inside <desc> elements.

BNE

According to the Biblioteca Nacional de España requirements the element to be translated at the first stage is the element <dc:description>.

In other words all resource related descriptive information. BNE will provide MARC (not DC) metadata so we can suppose following projection to TEI P5:

The supposed projection to TEI P5 (BNE):[2]



[2] See bigger version in Annex 1

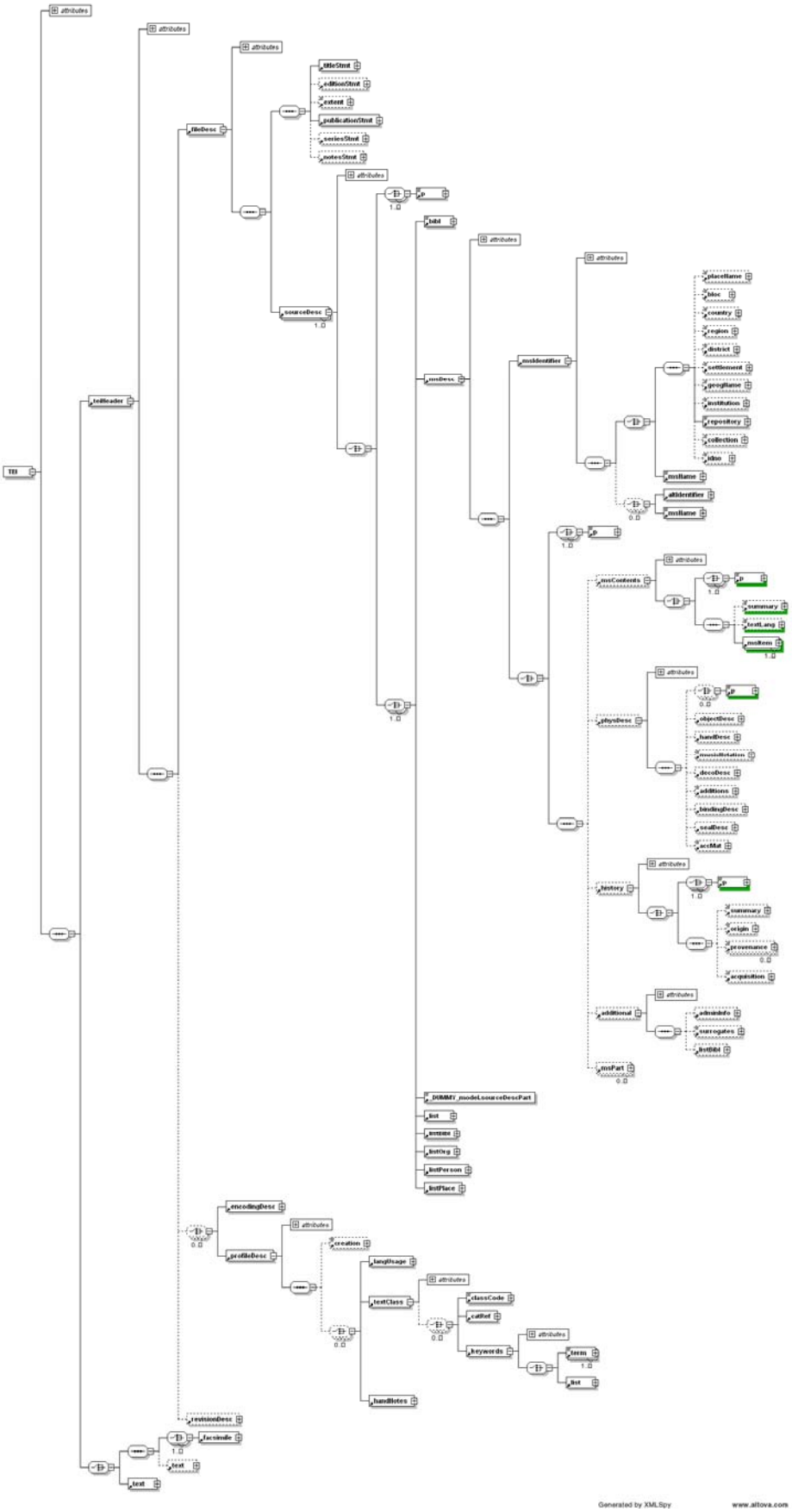
DSP

According to DSP only the following elements should be translated:

```
<title>
<textLang>
<note>
```

This information is interpreted as any note and general information related to the resource, textLang of the resource and any title used within the resource (as the head element will not be used within ENRICH TEI schema therefore below marked is the msContents/summary or msContent/msItem where most probaly the title information will be targeted).

The supposed projection to TEI P5 (DSP):[3]



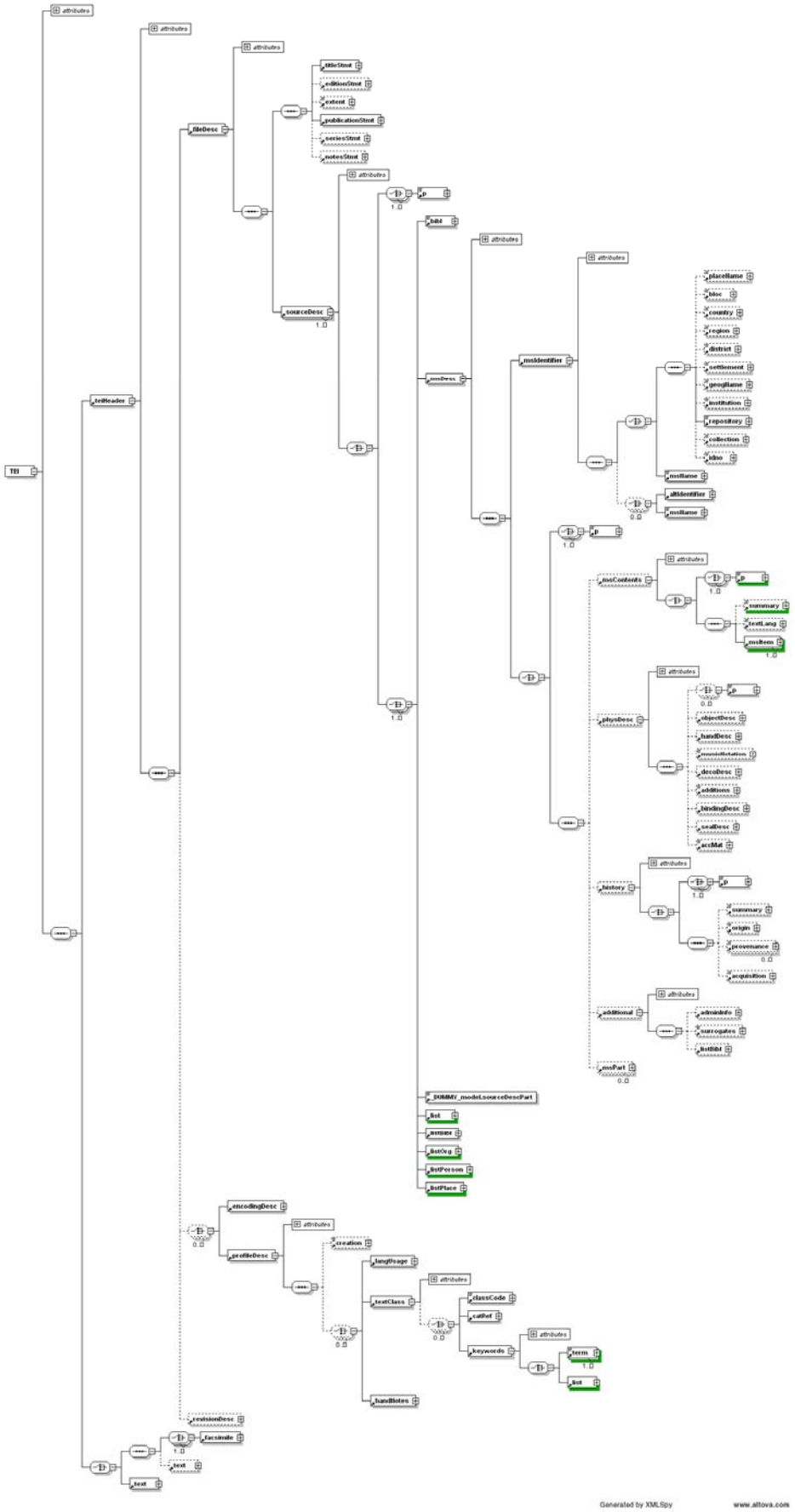
[3] See bigger version in Annex 1

ULW

Based on Wroclaw University Library (WUL) translation will be needed for the following elements

-main title
-title of the chapter/article
-keywords
-description of the document's content
-description of the content of the part (chapter)
-coverage: geographical place or region
-coverage: period, time
-details about the author, contributor

The supposed projection to TEI P5 (ULW):[4]



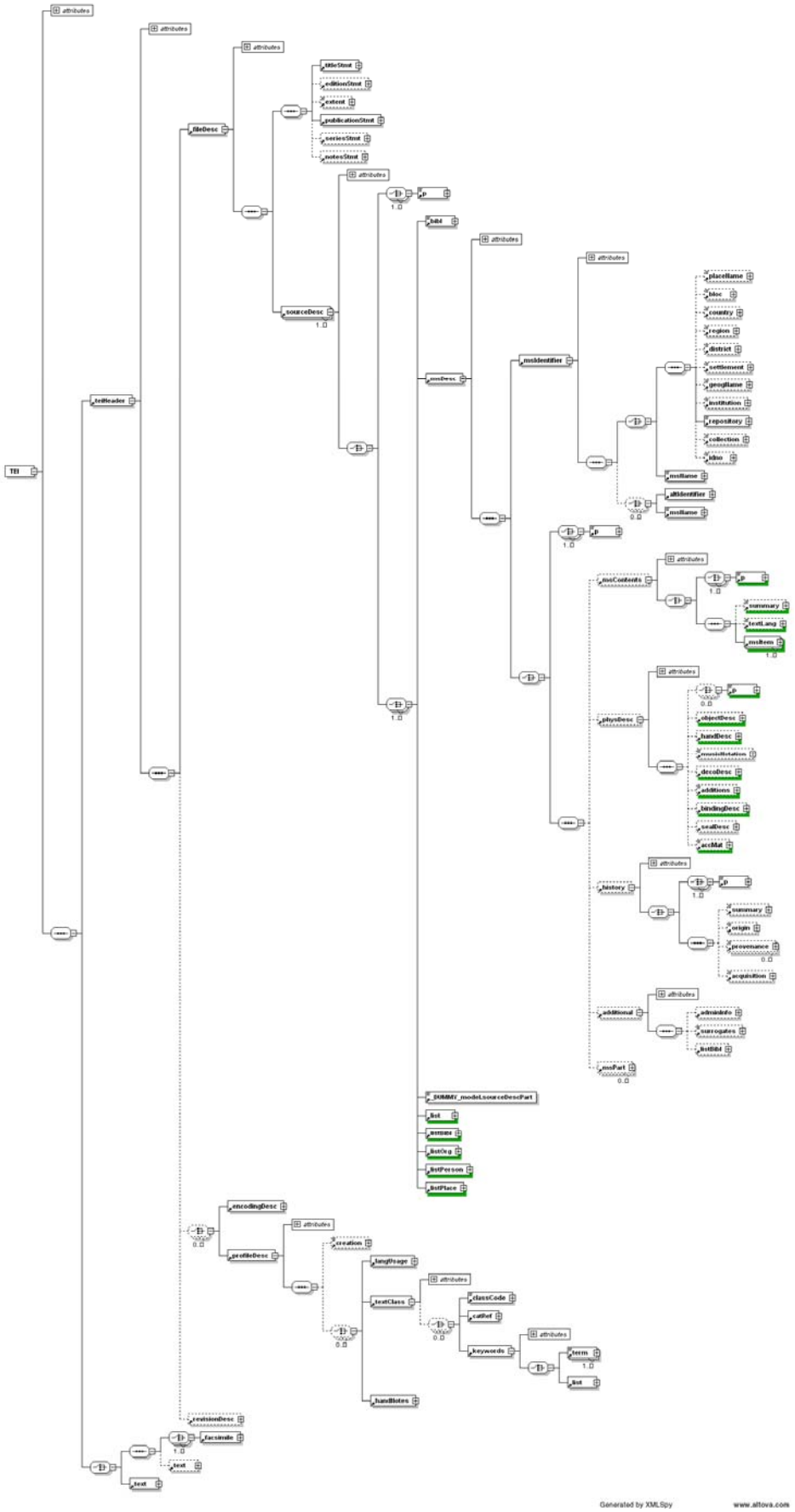
[4] See bigger version in Annex 1

BNCF

The following elements have to be translated:
df t = "3xx"
df t = "6xx"
where 3xx indicates all the tags beginning with 3 and 6xx indicates all the tags beginning with 6.

In other words mainly so called note fields and classification fields which will target the list elements (additional diagram) and the content of XXXX (tady resFields etc).

The supposed projection to TEI P5 (BNCF):[5]



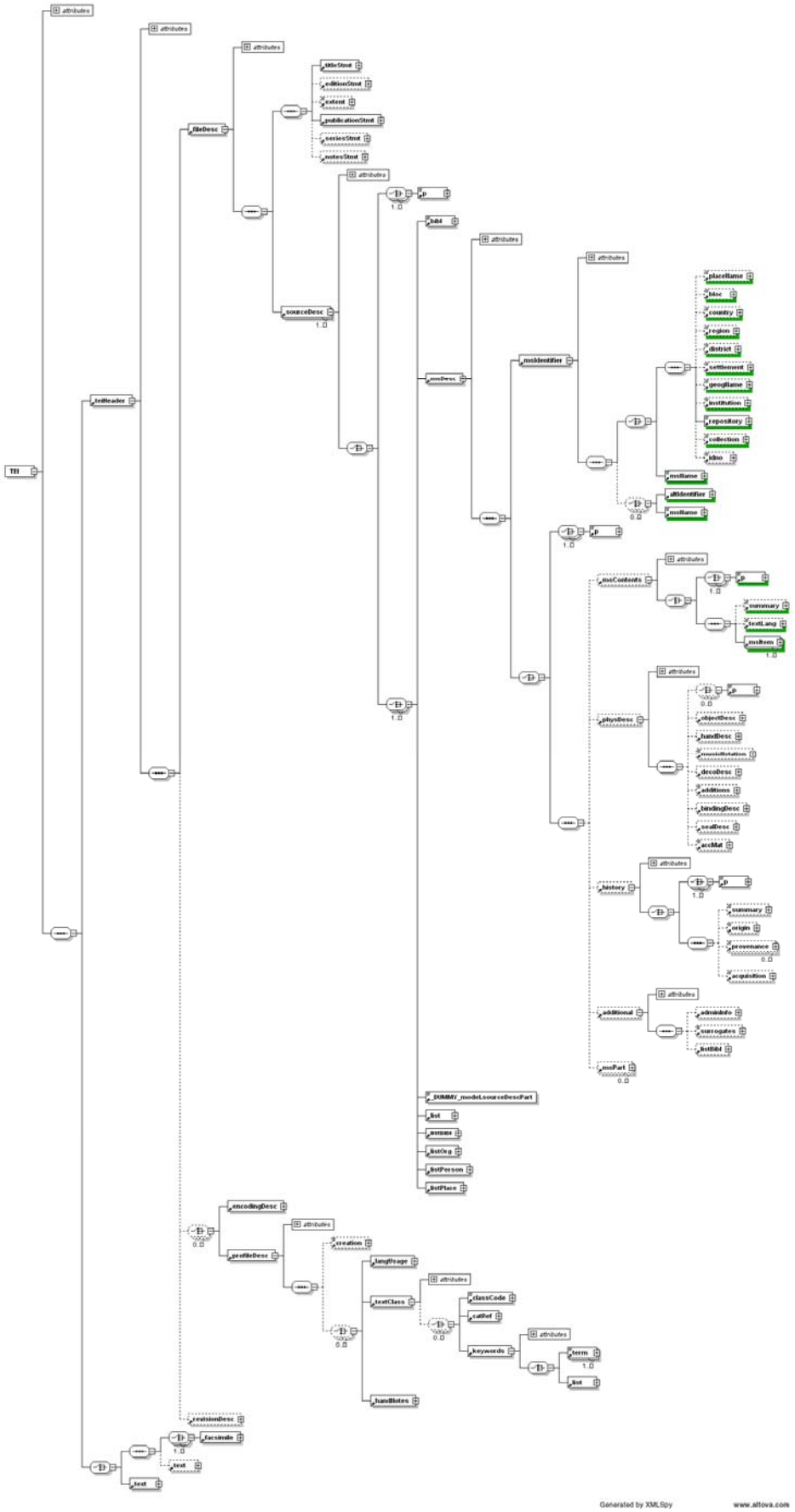
[5] See bigger version in Annex 1

UZK

According to their requirements the following element need to be excluded from translation:

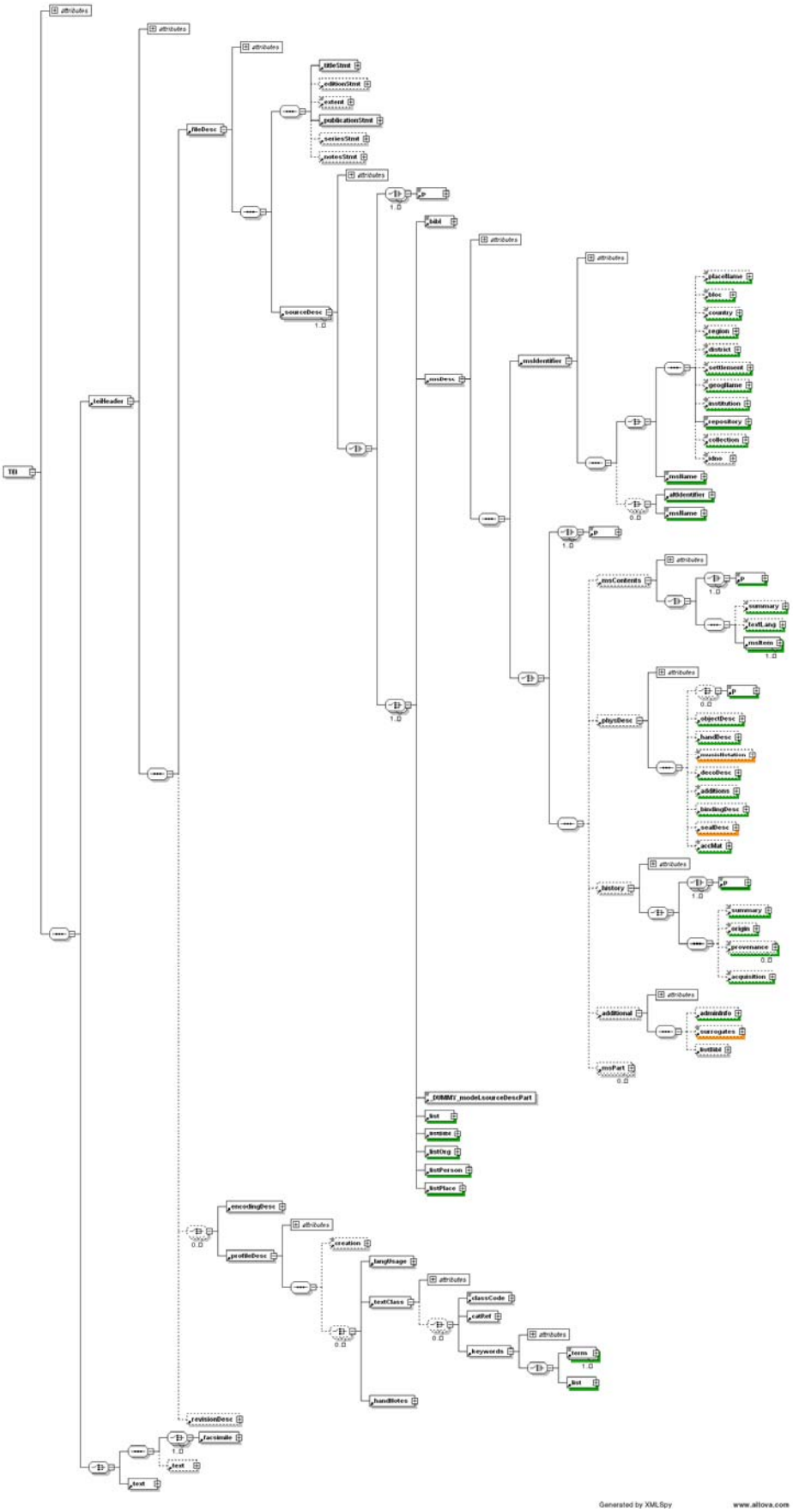
- //bibl
- //explicit
- //incipit
- //locus
- //msContents
- //msIdentifier
- //ref
- //title

The supposed projection to TEI P5 (UZK):[6]



[6] See bigger version in Annex 1

By the disjunction of the set of TEI P5 elements for translation we may get the expected set of elements for translation:[7]



[7] See bigger version in Annex 1

The orange elements are added artifficialy as these represent the same type of information as similar green marked (and therefore should be included into the translation in the wiev to general further sources processing). This diagram gives quite good idea of what will be translated.

Reffering now tho the set of information as described in the chapter “1. Metadata content” the full **resourceInfo** set of information should be translated.

- **resourceInfo**
  - **resIdentifier**
  - **resDesc**
    - **resContents**
    - **resResp**
    - **resPhysDesc**
    - **resHistory**
  - **resAdditional**
  - **resKeywords**
- **digidocInfo**
  - **digidocStruct**
  - **digidocTechDesc**
- **adminInfo**
  - **adminIdentification**
  - **adminResp**
  - **adminRights**

In other words: full bibliographic description of the resource in modern languages should be translated, no Content Partner expressed the need to translate the **adminInfo** and **digidocInfo** set of information.

III Conclusions

The Corpus analysis has been based on current formalisms used by the different institutions as well as their requirements concerning the element(s) to be translated. All partners agree on the translation of the **resourceInfo** set of elements.

It is proposed to indicate the language of the contents of the particular element, to be used in both resource discovery and in filtering retrieval results

In the case of the historical documents to be made accessible via the ENRICH project it is of major importance to record the titles in several languages:

-the original language of the document. This will be in most cases in an old version of the current language: old Hungarian, old German, old French, using words, which disappeared by now and an orthography quite different from the modern one,

-the modern transcription of the title, with appropriate modern terms and orthography, for retrieval purposes

- the translation in English for the purposes of a unique search throughout the data base.

SYSTRAN exploited those conclusions in order to start developping the first version of the SYSTRAN Translation Stylesheet (Task 6.2) that is used to drive a complex external process with parameters: a Machine Translation system. This mechanism binds the deep structure of the source document to the translation engine options rendering a correlation between the document syntax and its underlying linguistic structure. Conversely, one can enrich the document structure with linguistic information in order to improve the quality of machine translation results. In an optimal workflow, the author and the translation engine interact with a feedback - structure enrichment cycle supported by the users feedback platform http://enrich.systran.fr/enrich/feedbackForm.jsp. Based on the content partners feedback, SYSTRAN will adjust the SYSTRAN Translation Stylesheet according to the common formalism that will be adopted at the end of the project as well as the partners and users final requirements concerning the elements to be translated.

http://enrich.manuscriptorium.com/files/ENRICH_WP8_D_8_4_leaflet.pdf

Grant Agreement Number **ECP 2006 DILI 510049**

ENRICH

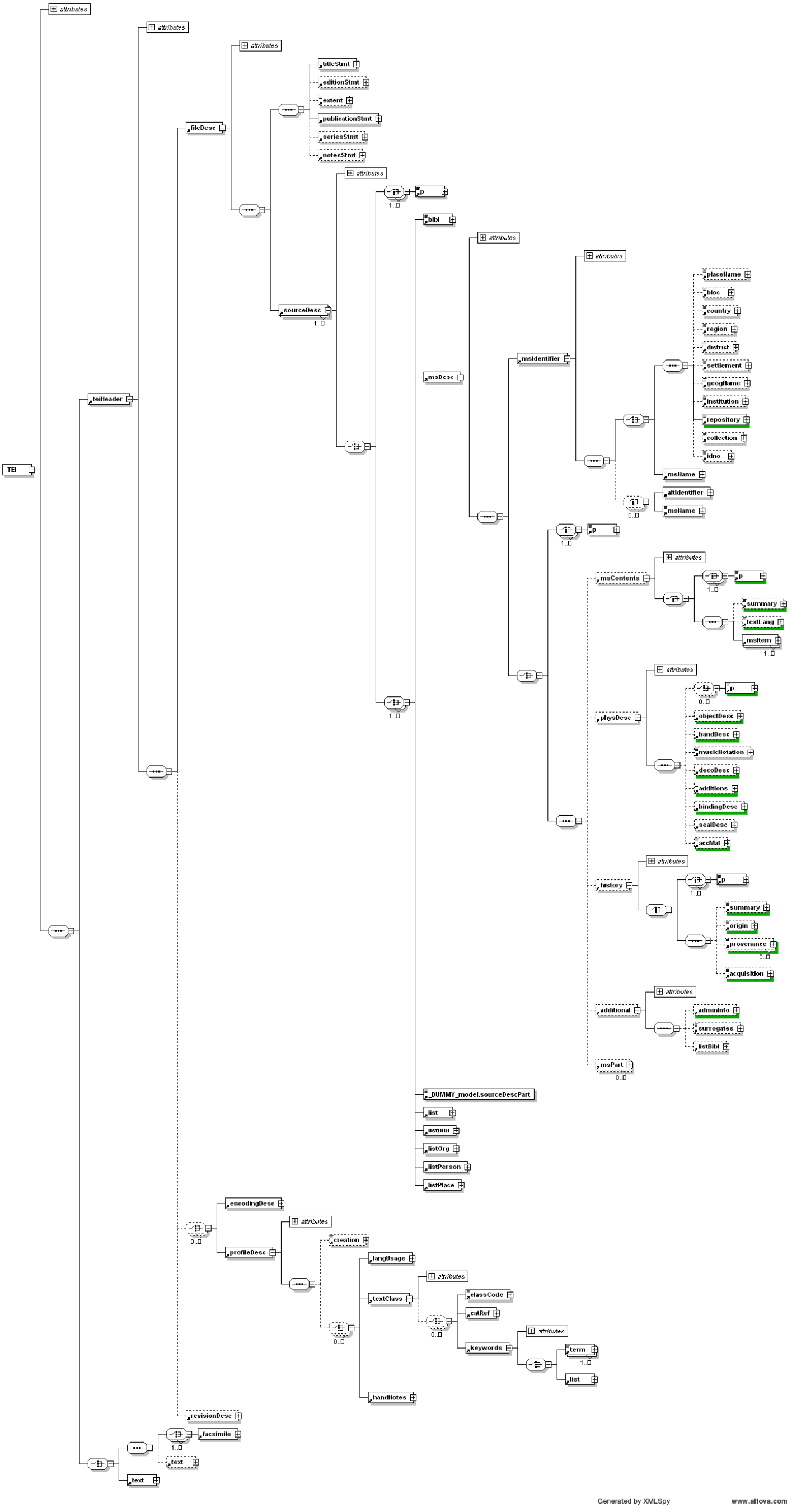
ENRICH Corpus Analysis report
Annex 1:
Supposed projections to TEI P5 for ENRICH partners

Deliverable number	***D6.1 Annex 1***
Dissemination level	*Public*
Delivery date	*26 September 2008*
Status	*Final*
Author(s)	*Elsa Zipstein*

e*Content*plus

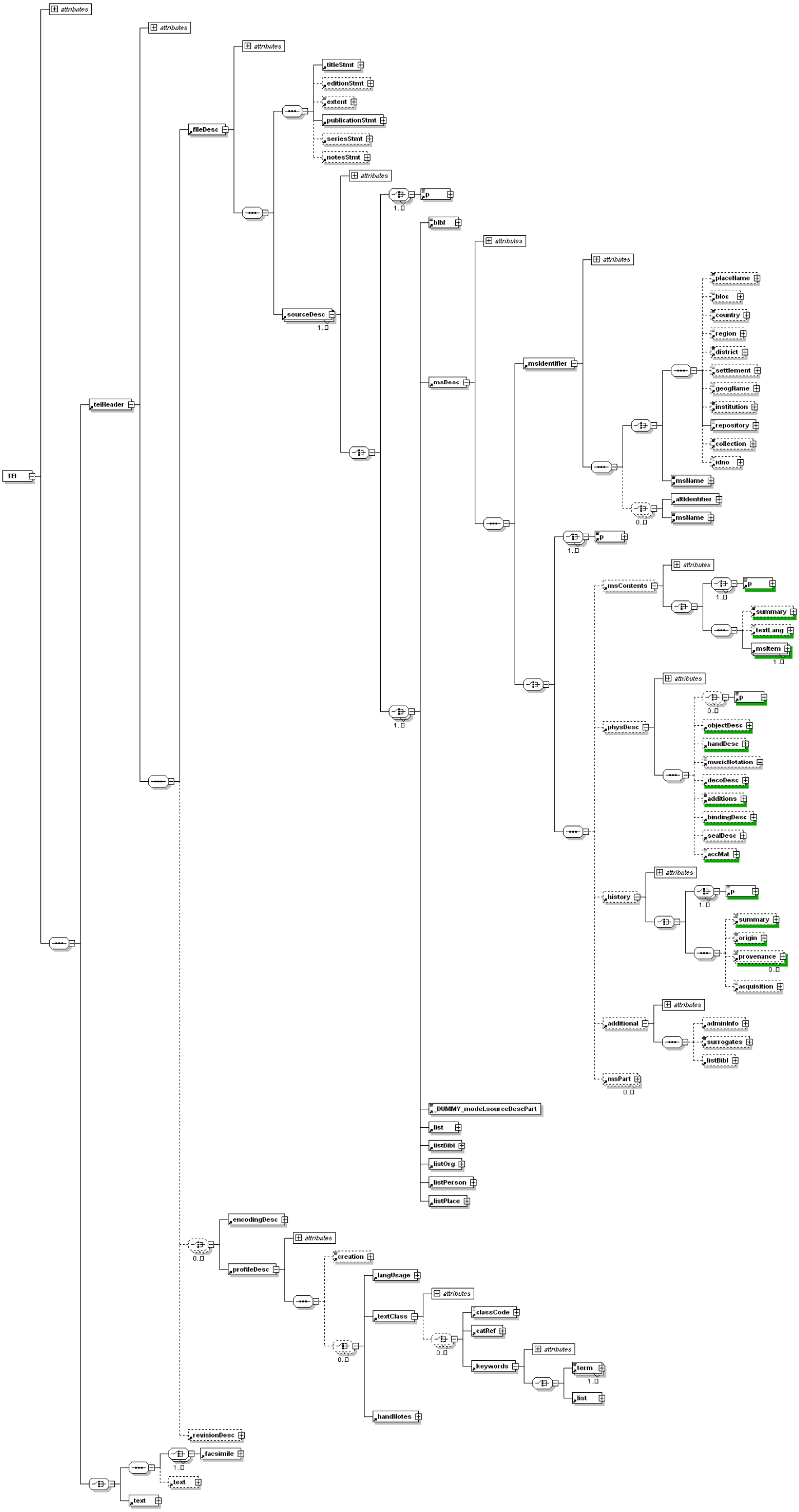
This project is funded under the *e*Content*plus* programme,
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

The supposed projection to TEI P5 (NKP):



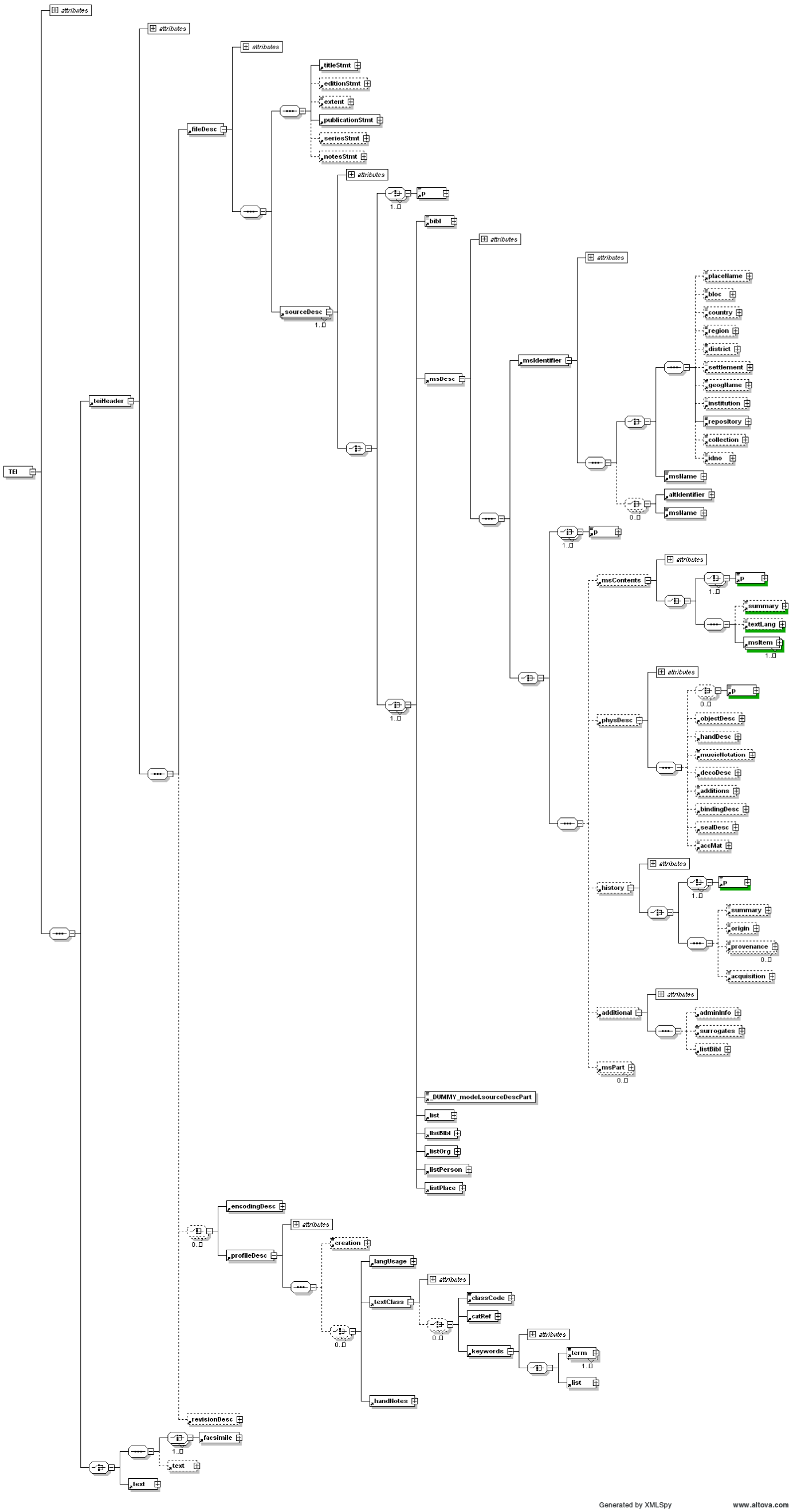
Annex 1: Supposed projections to TEI P5 for ENRICH partners

The supposed projection to TEI P5 (BNE):



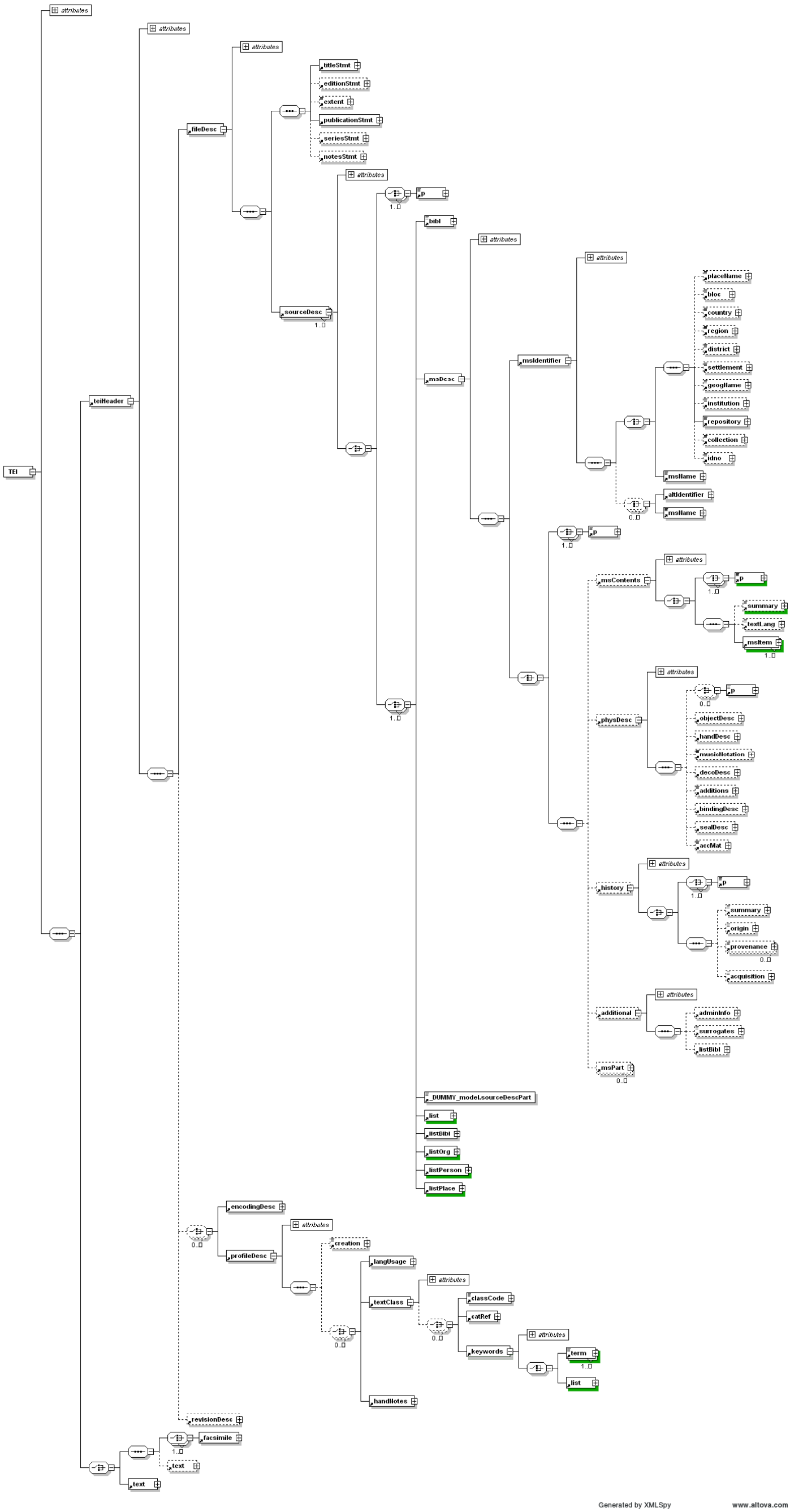
Annex 1: Supposed projections to TEI P5 for ENRICH partners

The supposed projection to TEI P5 (DSP):



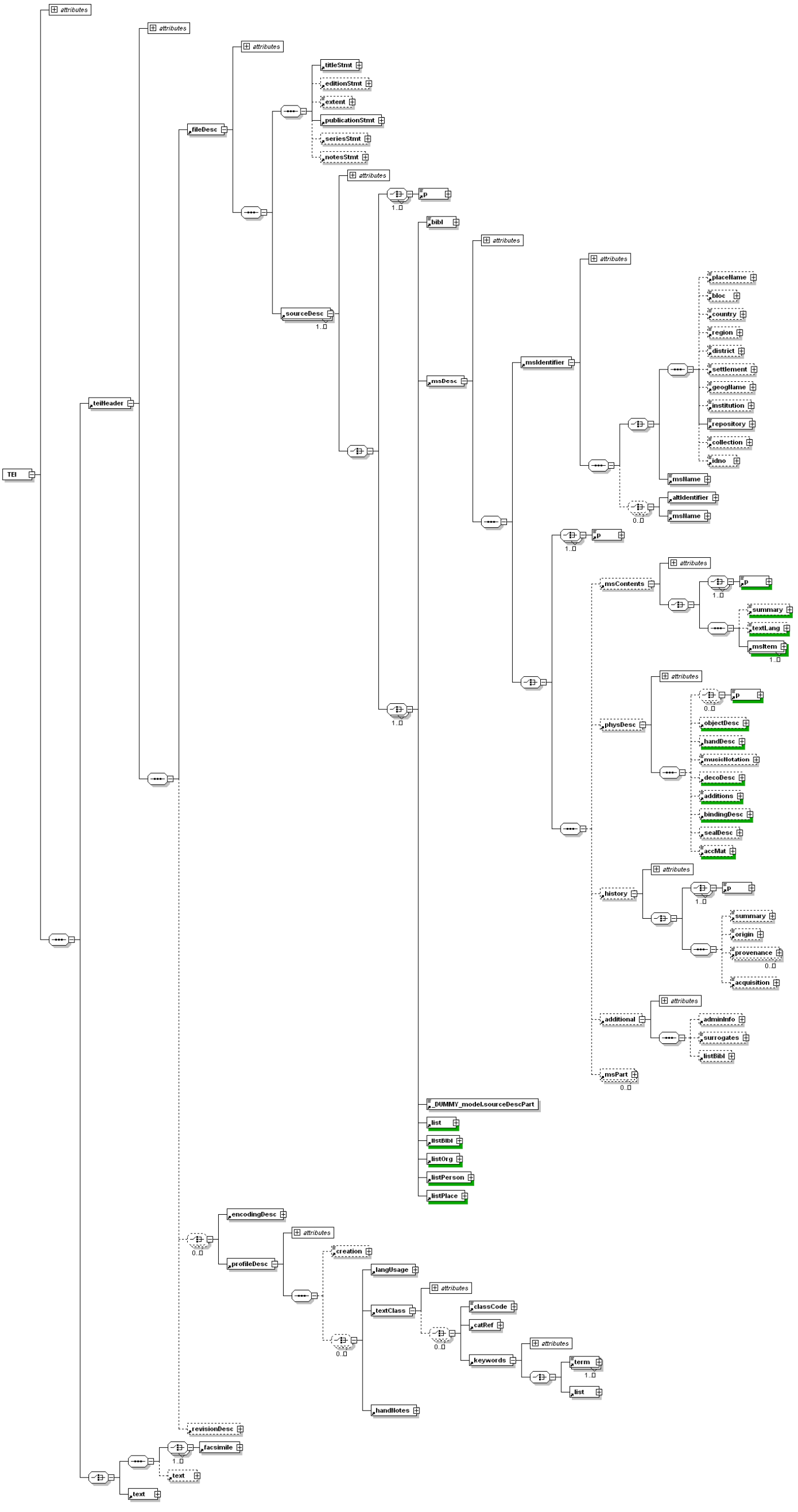
Annex 1: Supposed projections to TEI P5 for ENRICH partners

The supposed projection to TEI P5 (ULW):



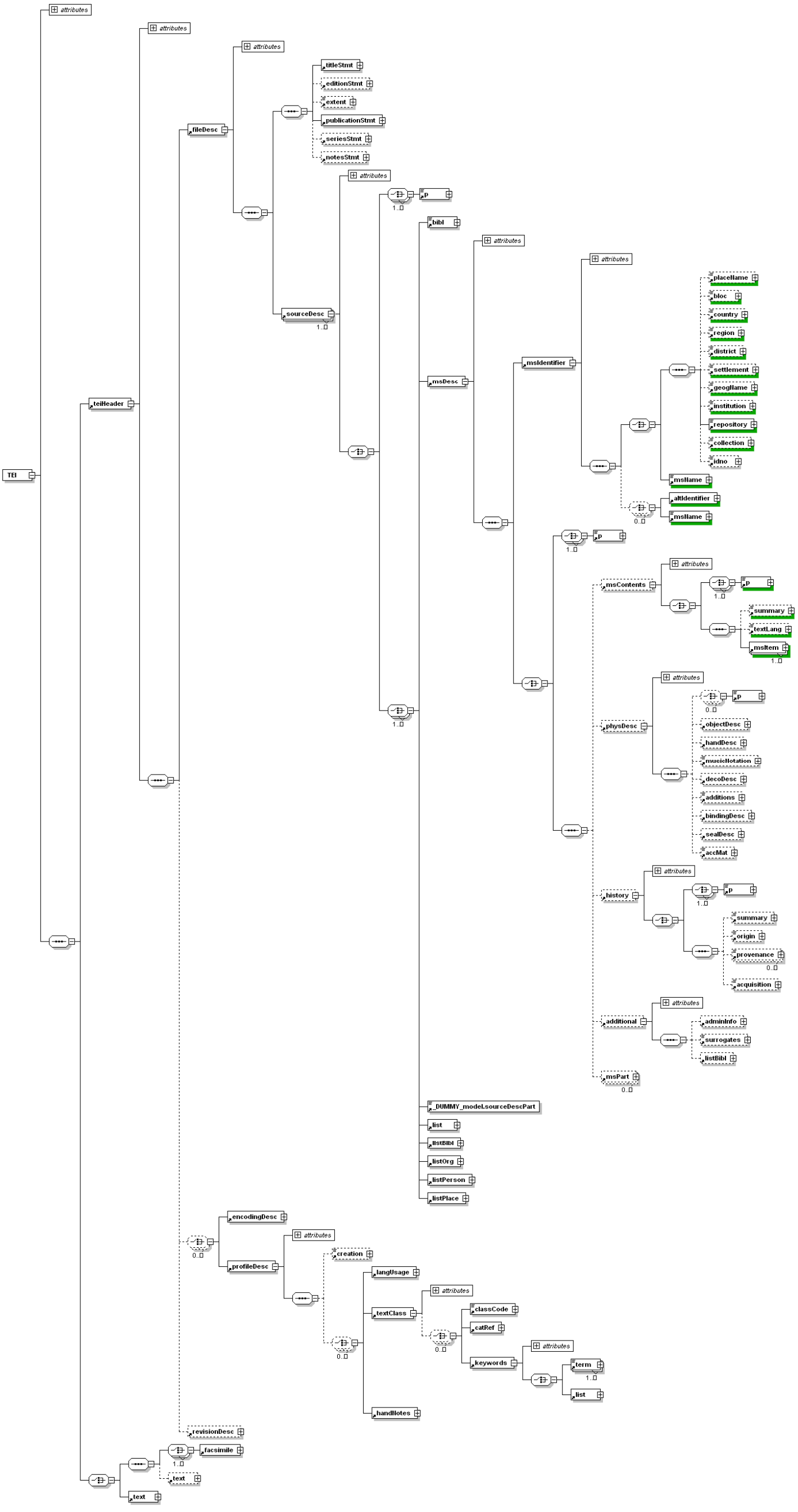
Annex 1: Supposed projections to TEI P5 for ENRICH partners

The supposed projection to TEI P5 (BNCF):



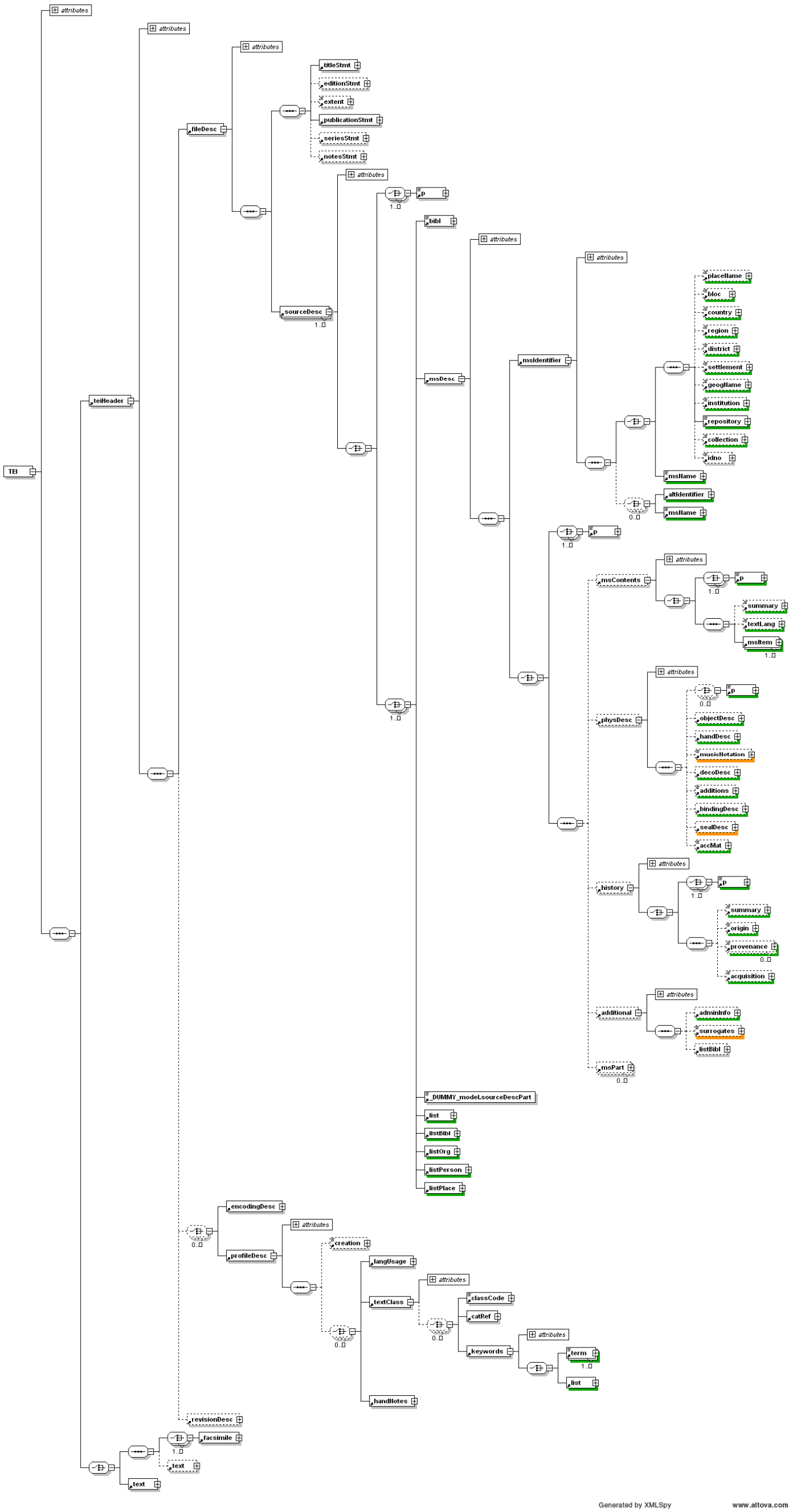
Annex 1: Supposed projections to TEI P5 for ENRICH partners

The supposed projection to TEI P5 (UZK):



Annex 1: Supposed projections to TEI P5 for ENRICH partners

By the disjunction of the set of TEI P5 elements for translation we may get the expected set of elements for translation:



Annex 1: Supposed projections to TEI P5 for ENRICH partners