The Text Encoding Initiative as a tool for manuscript description

outline and practical application

Galla Topalian[1] Jean-Baptiste Camps[2]

[1]Observatoire de Paris – [2]École des chartes | PSL

Shaping a European scientific scene: Alfonsine astronomy

ALFA is an ERC funded project for 60 month

(Consolidator grant 2016 agreement n° 723085)

ALFONSINE ASTRONOMY

Some questions

Given the scientific goals of the ALFA project, what would be the more suitable way to express manuscript descriptions

- in terms of codicological relevancy ?
- in terms of format and implementation?
- in terms of data life cycle,
  from production through exploitation to curation?

Outline

ALFA: goals and methodology

Goals

Retrace the development of the corpus of Alfonsine texts from its origin in the second half of the 13th century to the end of the 15th century by following, on the manuscript level, the milieus fostering it.

Questions of method

- Which conceptual framework for manuscript description?
- How to produce descriptions?
- What does it imply for the data?

Content & Container

Content: scientific aspects of manuscript description; data model;

Container: implementation and format.

Content models

- Very structured (e.g. database...);
- loosely structured and flexible (e.g. text-based description);
- both?

Containers

- (relational) database;
- XML document;
- spreadsheet (i.e. Excel, Calc...);
- text document.

What is the content model of a traditional manuscript description?

28843. In Latin, on parchment: written in more than one hand of the 13th cent. in England: 7¼ × 5⅝ in., i + 55 leaves, in double columns: with a few coloured capitals.

'Hic incipit Bruitus Anglie,' the De origine et gestis Regum Angliae of Geoffrey of Monmouth (Galfridus Monumetensis): *beg.* 'Cum mecum multa & de multis.'

On fol. 54ᵛ very faint is 'Iste liber est fratris guillelmi de buria de . . . Roberti ordinis fratrum Pred[icatorum],' 14th cent. (?): 'hanauilla' is written at the foot of the page (15th cent.). Bought from the rev. W. D. Macray on March 17, 1863, for £1 10*s*.

Now MS. **Add. A. 61.**

Figure 10.1. Entry for Bodleian MS. Add. A. 61 in Madan et al. 1895-1953

What is the content model of a traditional manuscript description?

28843. In Latin, on parchment: written in more than one hand of the 13th cent. in England: 7¼ × 5⅝ in., i + 55 leaves, in double columns: with a few coloured capitals.

'Hic incipit Bruitus Anglie,' the De origine et gestis Regum Angliae of Geoffrey of Monmouth (Galfridus Monumetensis): *beg.* 'Cum mecum multa & de multis.'

On fol. 54ᵛ very faint is 'Iste liber est fratris guillelmi de buria de . . . Roberti ordinis fratrum Pred[icatorum],' 14th cent. (?): 'hanauilla' is written at the foot of the page (15th cent.). Bought from the rev. W. D. Macray on March 17, 1863, for £1 10*s*.

Now MS. **Add. A. 61.**

Figure 10.1. Entry for Bodleian MS. Add. A. 61 in Madan et al. 1895-1953

Manuscript descriptions, both

- texts (beginning, end, reading order…);
- structured data (fields, controlled values, query-ready).

Prerequisites

The format must be chosen according to:

- the scientific purpose of the project;
- the project management (e.g.: partnership, reuse of other project method or result...);
- the desired workflow (from data creation to analysis, publication, reuse and curation).

Plan

Norms or standards for the description medieval manuscripts

In France, there is no specific and unified norm for the description of medieval manuscripts, particularly in the context of research.

Research-oriented recommendations

DFG Deutsche Forschungsgemeinschaft, *Richtlinien Handschriftenkatalogisierung* (Bonn-Bad Godesberg, 1992), `http://bilder.manuscripta-mediaevalia.de/hs/kataloge/HSKRICH.htm`.

IRHT Institut de recherche et d'histoire des textes, *Guide pour l'élaboration d'une notice de manuscrit*, 1977.

Bibliographic resources

- Paul Géhin, ed., *Lire le manuscrit médiéval: observer et décrire*, Paris, 2005.

Implementations

EAD (Encoded Archival Description)

- Format based on the ISAD(G) norm for the description of archives;
- sometimes used also for collections of manuscripts;
- used since 2002 by the *Bibliothèque nationale de France* and the *Catalogue général des manuscrits*.

Will be presented tomorrow!

TEI msDescription module

- conceived with Medieval Western European *codices* in mind;
- through several research projects involving European national or university libraries and research centres;
- in use in some national platforms (*E-Codices*), libraries (BAV,...) or research institutions (IRHT).

Plan

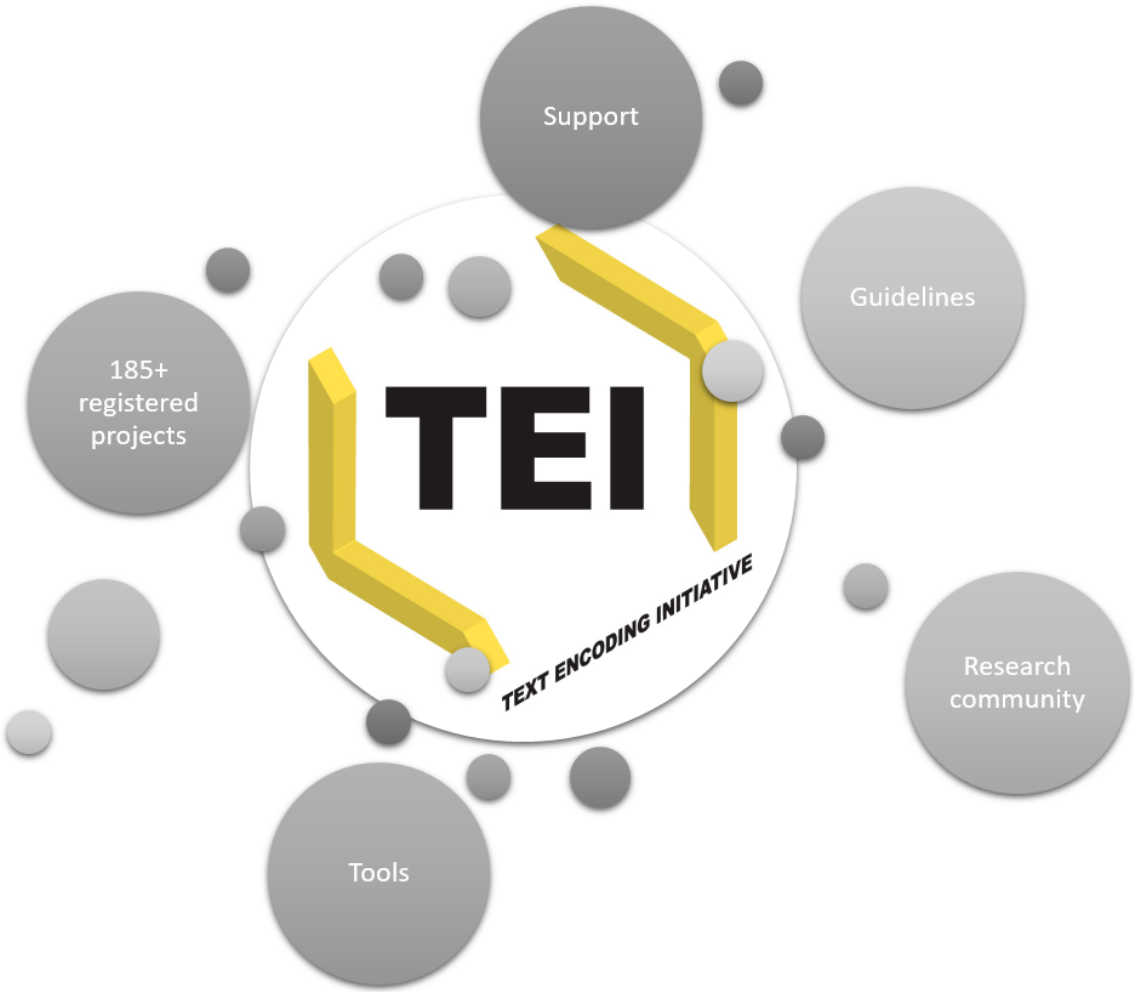
TEI: a conceptual framework created by researchers, for researchers


```
<TEI>
  <teiHeader>
    <fileDesc/>
  </teiHeader>
  <text>
    <body>
      <p>Salve!</p>
    </body>
  </text>
</TEI>
```

Important facts

- **1987**: Poughkeepsie meeting and establishment of the *Text Encoding Initiative*;
- **Goal**: "provide a standard format for data interchange in humanities research".
- **1990** TEI Proposal 1 (TEI P1), *Guidelines for the Encoding and Interchange of Machine-Readable Texts*, dir. Michael Sperberg-McQueen and Lou Burnard;
- **2000**: creation of the TEI Consortium;
- **2007**-...: TEI P5.

The TEI galaxy



Master, Enrich & msDescription

1999-2001: Manuscript Access through Standards for Electronic Records

Goal: to create elements in TEI for manuscript description and to establish a European standard.

> *To establish a single interchange standard for (...) descriptions of manuscripts*

See Lou Burnard, Muriel Gougerot, Elizabeth Lalou, et Peter Robinson, *Towards a European Standard for Manuscript Description: the MASTER project*, 1999.

2007-2009: European Networking Resources and Information concerning Cultural Heritage

Goal of "reducing the number of choices and constraining the possible values"
Creation of the Manuscriptorium platform.

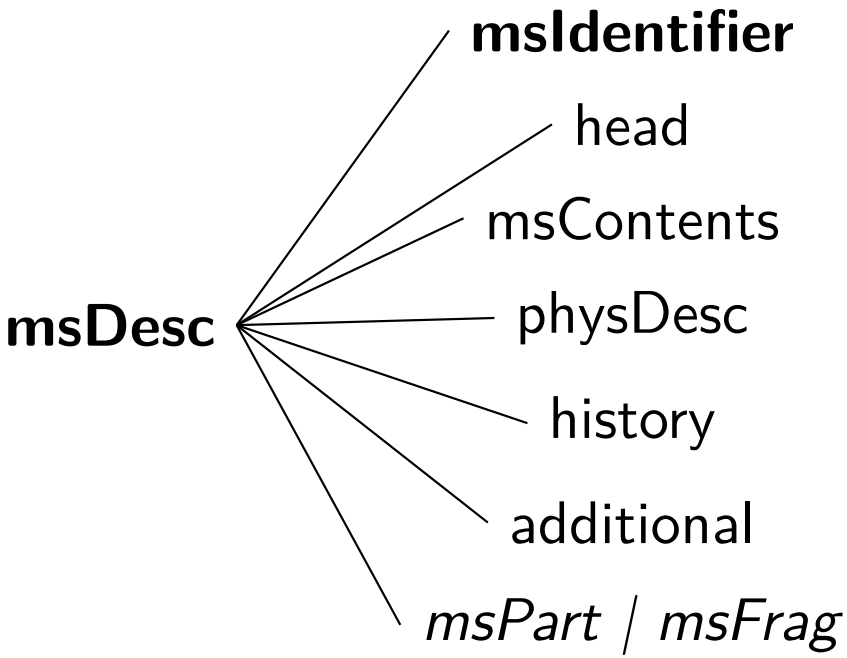
In a nutshell...

- A rich method created for research purpose;
- An active community providing methodology, support and re-usable data;
- An interoperable, persistent and open format

Plan

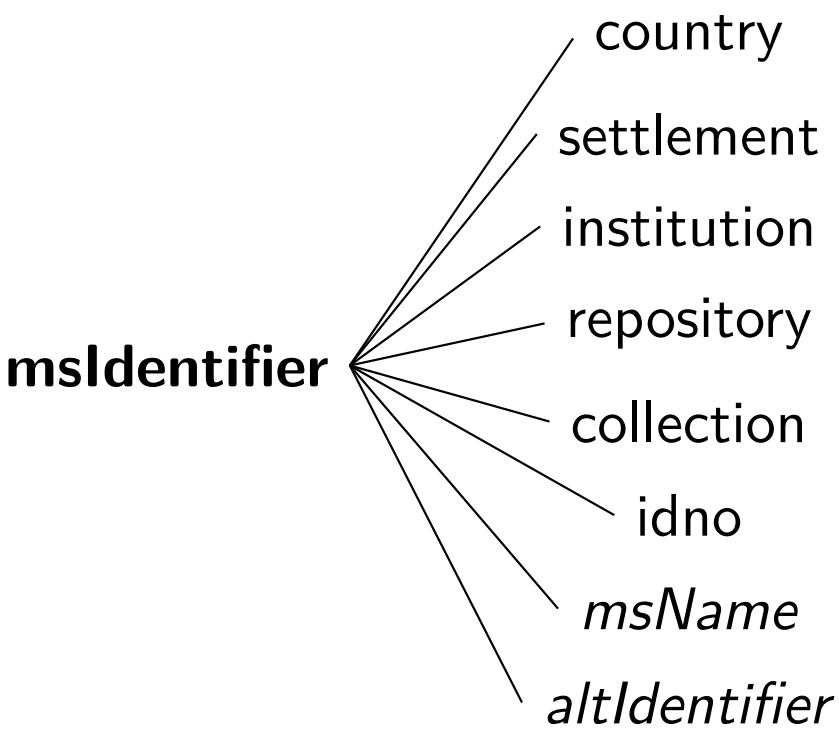
The msDesc element

Each description is contained in a msDesc element.



- shelfmark and related information;
- description title;
- (textual) contents;
- physical description;
- history of the ms.;
- other information;
- volume composed of several mss *or* hypothetical ms. of several fragments.

The ms. identifier (msIdentifier)

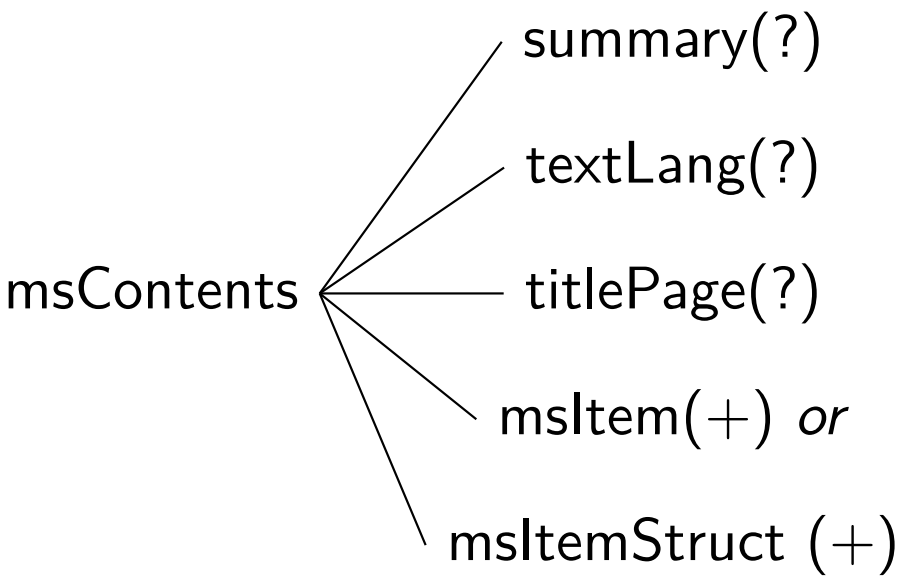


Informations related to current and old shelfmarks, in a structured way.

Problem

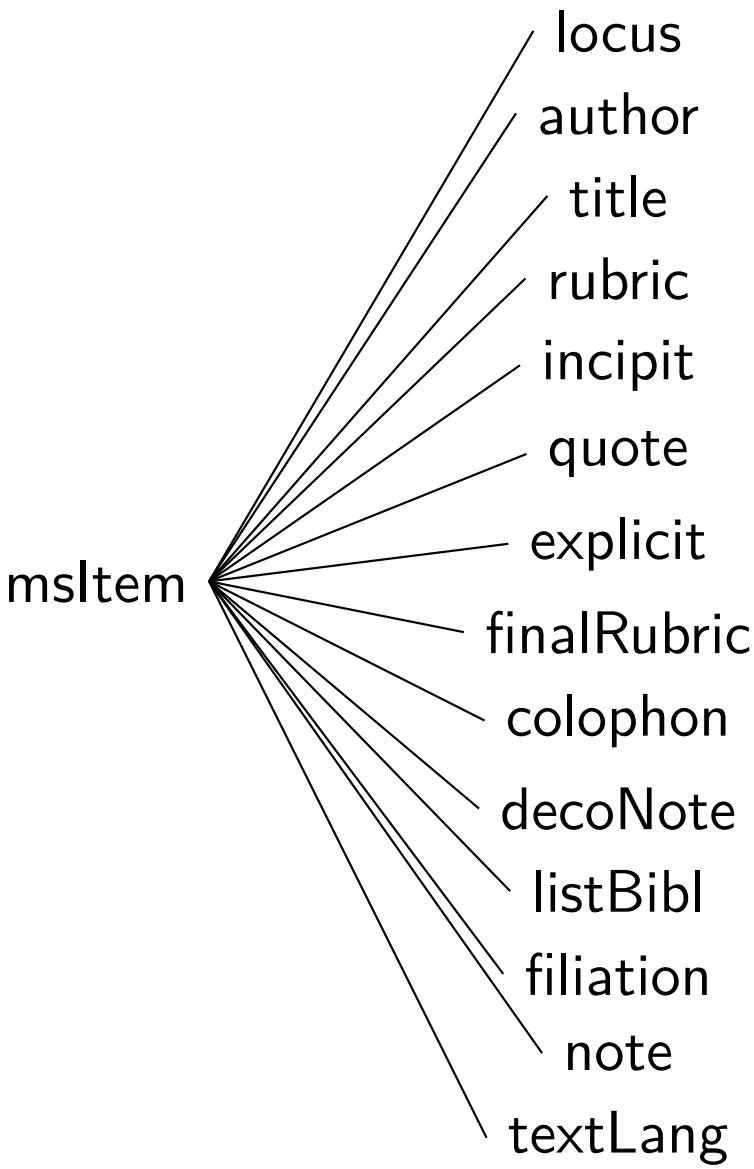
There is currently no standard identifiers for manuscripts.

The contents (msContents)



Either plain paragraphs *or* a more structured content.

A content item (msItem)



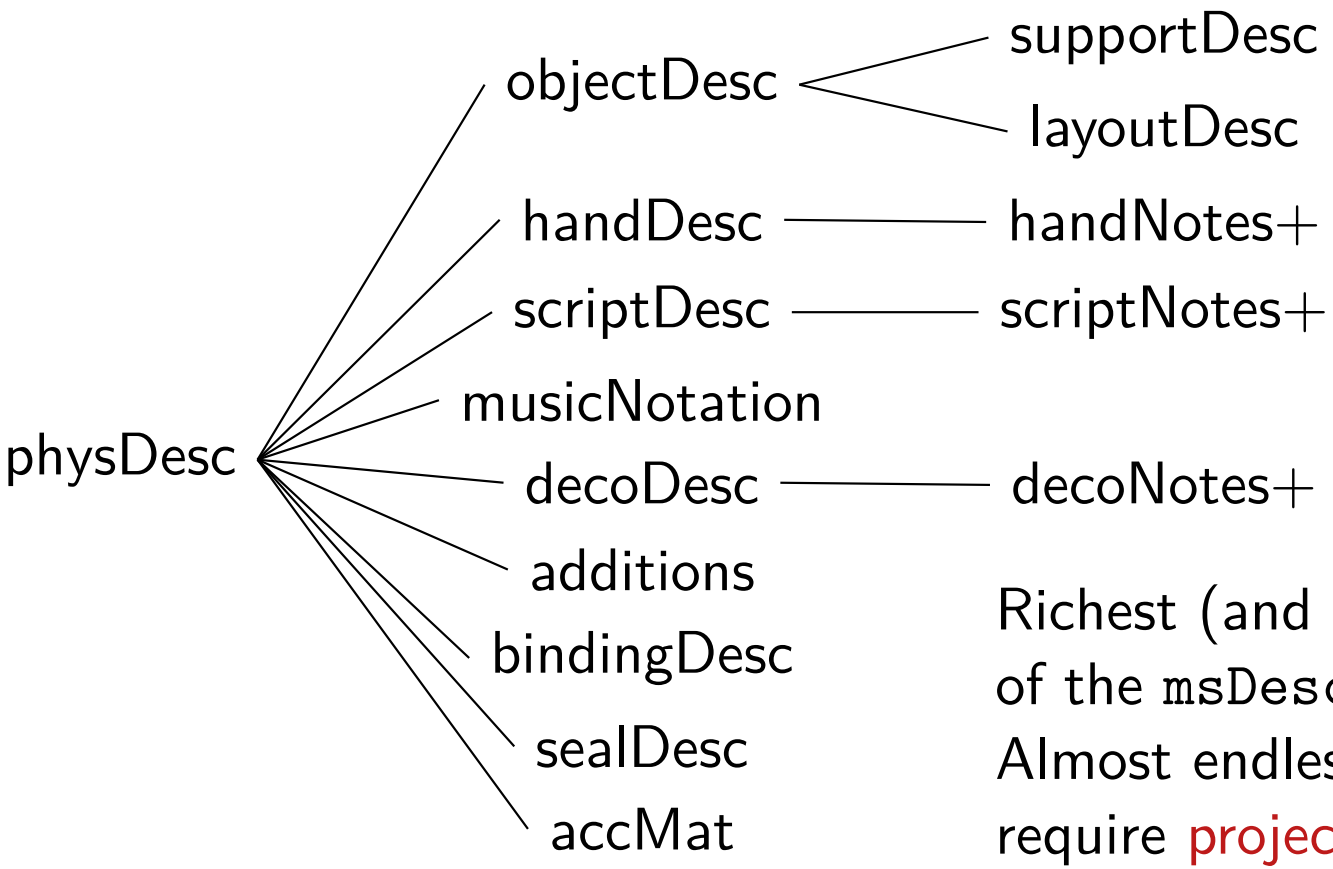
The description of a content item may be very precise, and include information regarding

- witness and filiation;
- paratext;
- scribal colophon...

It is possible to link information to authority files, e.g.:

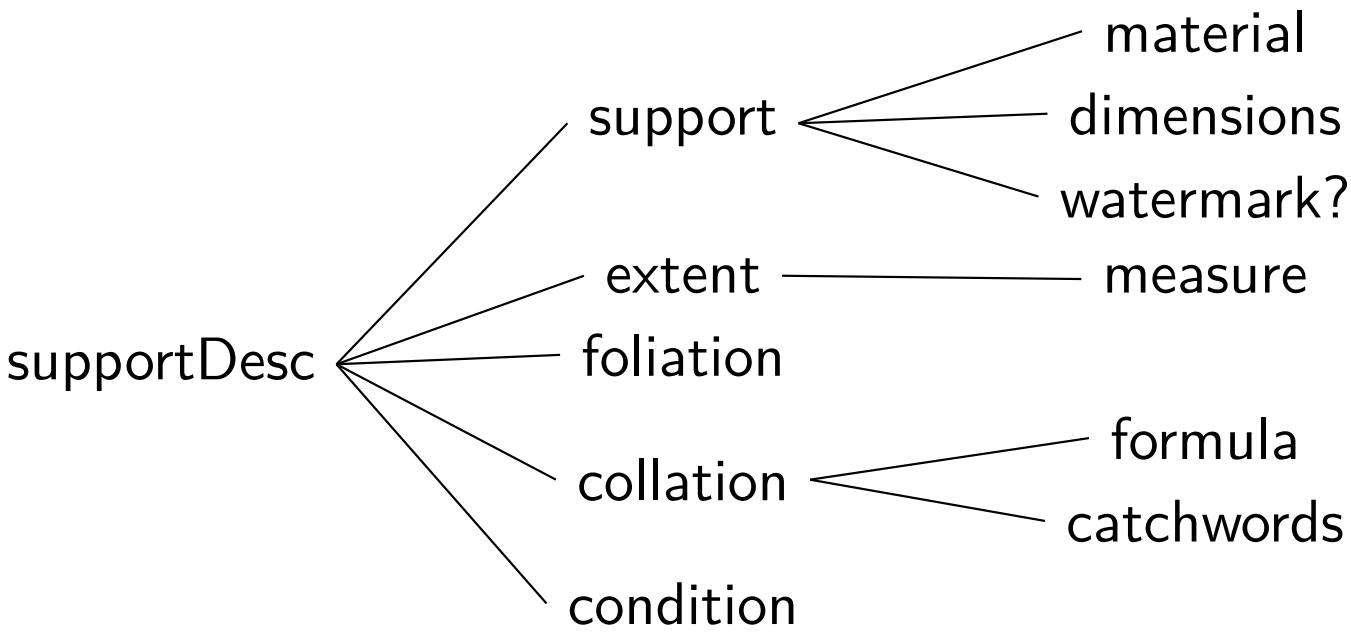
```
<title
  key="Otinel"
  ref="http://viaf.org/viaf/
  199768389"
  >Chanson d'Otinel</title>
```

The physical description (physDesc)

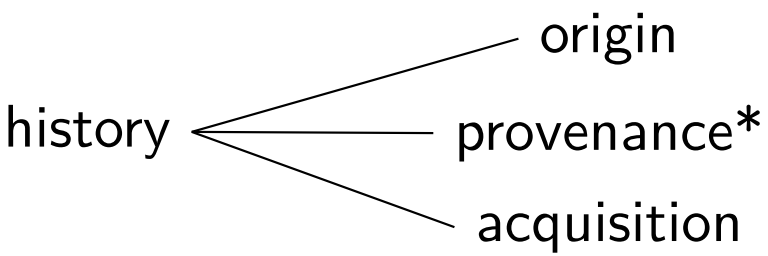


Richest (and most complex) part of the msDesc...
Almost endless possibilities that require project defined specifications to be usable...

An example: supportDesc



The history of the manuscript



The history of the ms., from its origin to its entry in its current collection.

May be further structured with `persName`, `placeName`, `origDate`, `origPlace`, and links to authority files.

Plan

Consistency and customisation

TEI Guidelines (chapter on customisation)

From the start, the TEI was intended to be used as a set of building blocks for creating a schema suitable for a particular project. This is in keeping with the TEI philosophy of providing a vocabulary for describing texts, not dictating precisely what those texts must contain or might have contained. This means that it is likely, not just possible, that you will want to have a tailored view of the TEI.

Burnard, 2015

Almost no-one needs everything defined by the TEI, yet every one of its elements is of use or interest to someone.

Customising, yes, but what?

- Level of granularity

msContents: short version

```
<msContents>
 <p>A collection of Lollard sermons</p>
</msContents>
```

msContents: slightly longer and more structured version

```
<msContents>
<msItem n="3">
  <locus from="211" to="222" scheme="#principale">fol. 211-222</locus>
  <title key="Otinel" ref="http://viaf.org/viaf/199768389"
  >Chanson d'Otinel</title>
  <incipit>Ki volt oïr cha<expan>n</expan>çu<expan>n</expan> de
  beau se<expan>m</expan>bla<expan>n</expan>t</incipit>
  <explicit>Ke si aida païens a traverser</explicit>
  <filiation>Témoin <idno>B</idno>, de 1908 vers, qui est le seul complet,
portant un texte voisin de M [...]</filiation>
  <decoNote>
    <p><term>Grandes initiales</term> de <measure unit="lines">4</measure>
    lignes de réglure au début du texte.</p>
    <p><term>Moyennes initiales</term> ... </p>
  </decoNote>
  <textLang>scripta anglo-normande, de la <origDate notBefore="1250"
  notAfter="1300" >deuxième moitié du XIIIe siècle</origDate></textLang>
</msItem>
</msContents>
```

Customising, yes, but what?

- Level of granularity
- **Scope of description**

Scope of description

TEI doesn't control the scope of the description:

> msIdentifier, is the only [component] which is mandatory; [...]. It is followed optionally by one or more head elements [...] and then either one or more paragraphs, marked up as a series of p elements, or one or more of the specialized elements msContents [...], physDesc [...], history [...], and additional [...]. These elements are all optional, but if used they must appear in the order given here.

Customising, yes, but what?

- Level of granularity
- Scope of description
- **Chosen vocabulary**

Chosen vocabulary

TEI is extremely loose in the attribute value. For consistency purpose, projects need to create custom vocabulary lists. E.g.: list of Enrich custom value for the script attribute:

- carolmin
- textualis
- cursiva
- hybrida
- humbook
- humcursiva
- other

Restraining and enriching the TEI

Collation formulas

A TEI customisation is always a restriction of the possibilities offered by TEI.
Yet, it can be somewhat more expressive by defining formal ways of representing some phenomena.

Collation formula, sample implementation: Chroust notation

```
<collation>
  <formula notation="Chroust"> I<hi rend="sup">2</hi>
  + (V-1)<hi rend="sup">11</hi>
  + 21 IV<hi rend="sup">169</hi>
  + (III-1)<hi rend="sup">184</hi>
  </formula>
</collation>
```

Collation formula, another possibility of the same

```
<collation>
  <measure type="quire" n="1">1-1</measure>
  <measure type="quire" n="2">5-4</measure>
  <measure type="quire" n="3">4-4</measure>
</collation>
```

Needing more?

- Creating new tags (not recommended) or asking the Consortium for new ones (if really needed);
- Using other namespaces: Music Encoding Initiative (for scores), MathML (for formulas), ... but also ComicsML (for comic books!).

Plan

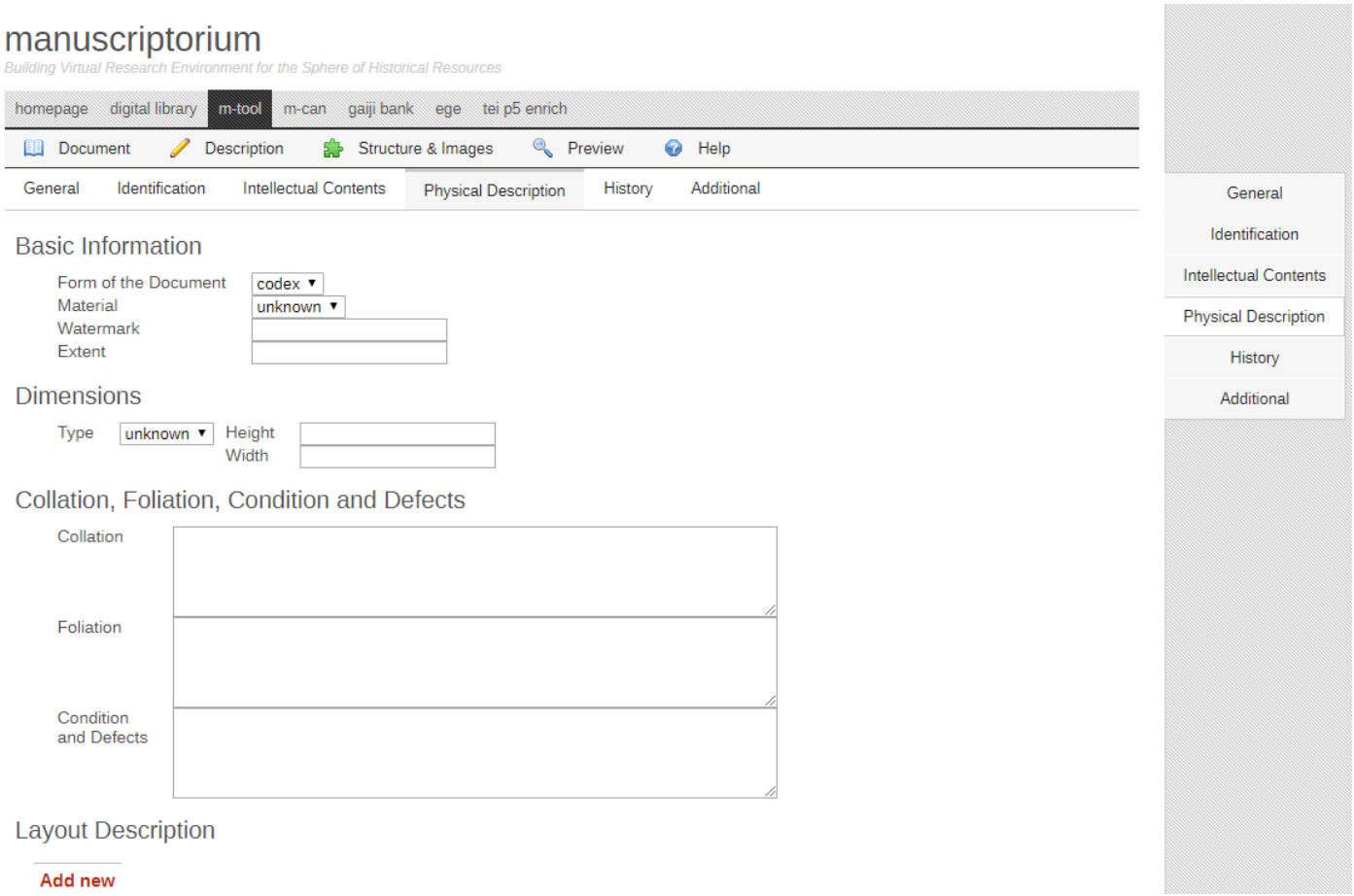
TEI from scratch

Two editors are notably used for editing TEI :

- Oxygen (alas, proprietary);
- XMLMind.

In both cases, the usage of these tools require some understanding of the code.

Production interface Explore m-tool



- Manuscriptorium project
- based on ENRICH specification

Conversion and edition of existing document

Through `xslt` transformations (and/or manual editing), it is possible to

TEI -> TEI ALFA

Get and convert to the project-specific TEI some descriptions in another TEI specification (e.g. from Manuscriptorium, E-Codices, etc.).

EAD,MarcXML -> TEI ALFA

Convert EAD or other XML documents to TEI (within the limits of possible mappings), and further enrich them.

Plan

Documents as databases

Indexing descriptions

XML structure makes it possible to use documents as databases and query their content.

Given the proper structure, one can retrieve all descriptions where

- a given person is mentioned in the `origin` section;
- the sum of page dimensions is inferior to n;
- the quires used are quaternions and the origin date is the XIVth century.

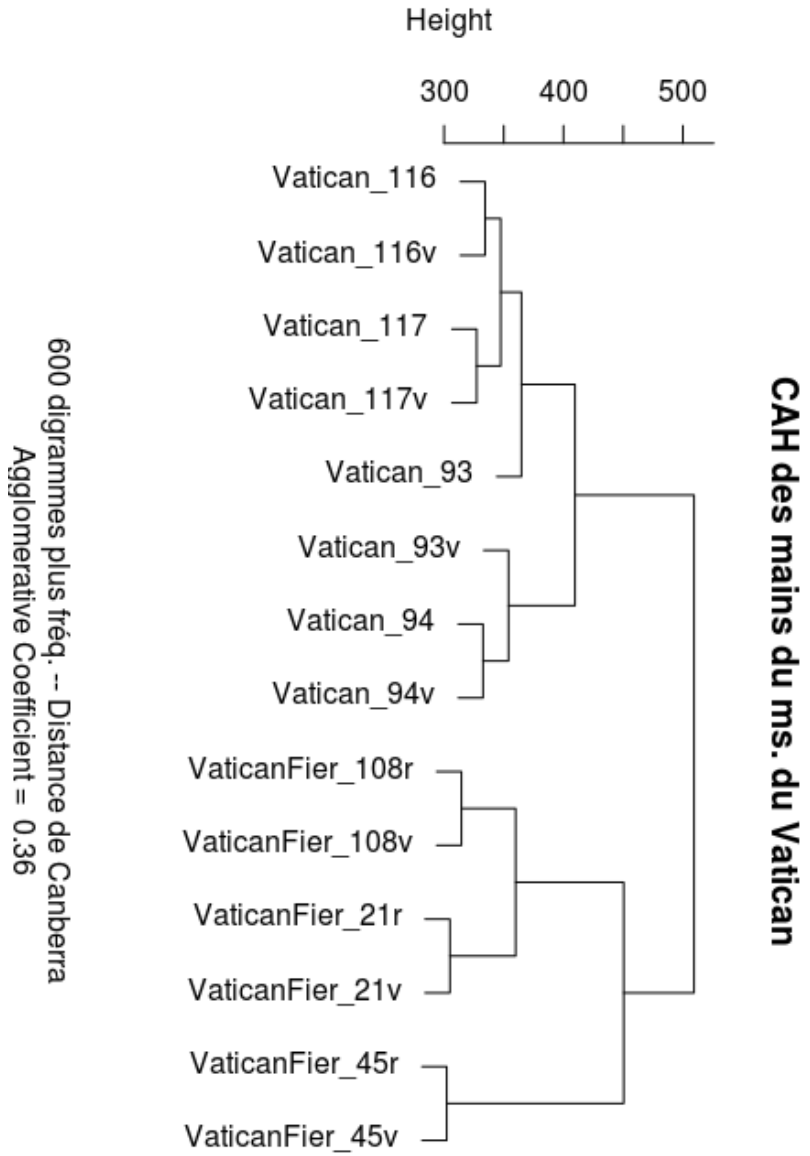
Crossing descriptions and transcriptions

TEI documents can also contain transcriptions or editions.

Indexing makes it possible to cross codicological with textual features, e.g. retrieve all documents where there is

- red painted initials and a given astronomical formula.

Data mining and quantitative codicology



TEI encoding makes it possible to

- extract some features from the data;
- convert it to csv for analysis (e.g. with the statistical software R).

Publishing the data

Stylesheet conversions

(Customisable) stylesheet already exist to convert TEI to:

- LaTeX for print publishing (or `odt`, `doc`...);
- `ePub` for e-readers and tablets;
- and of course HTML for online publishing.

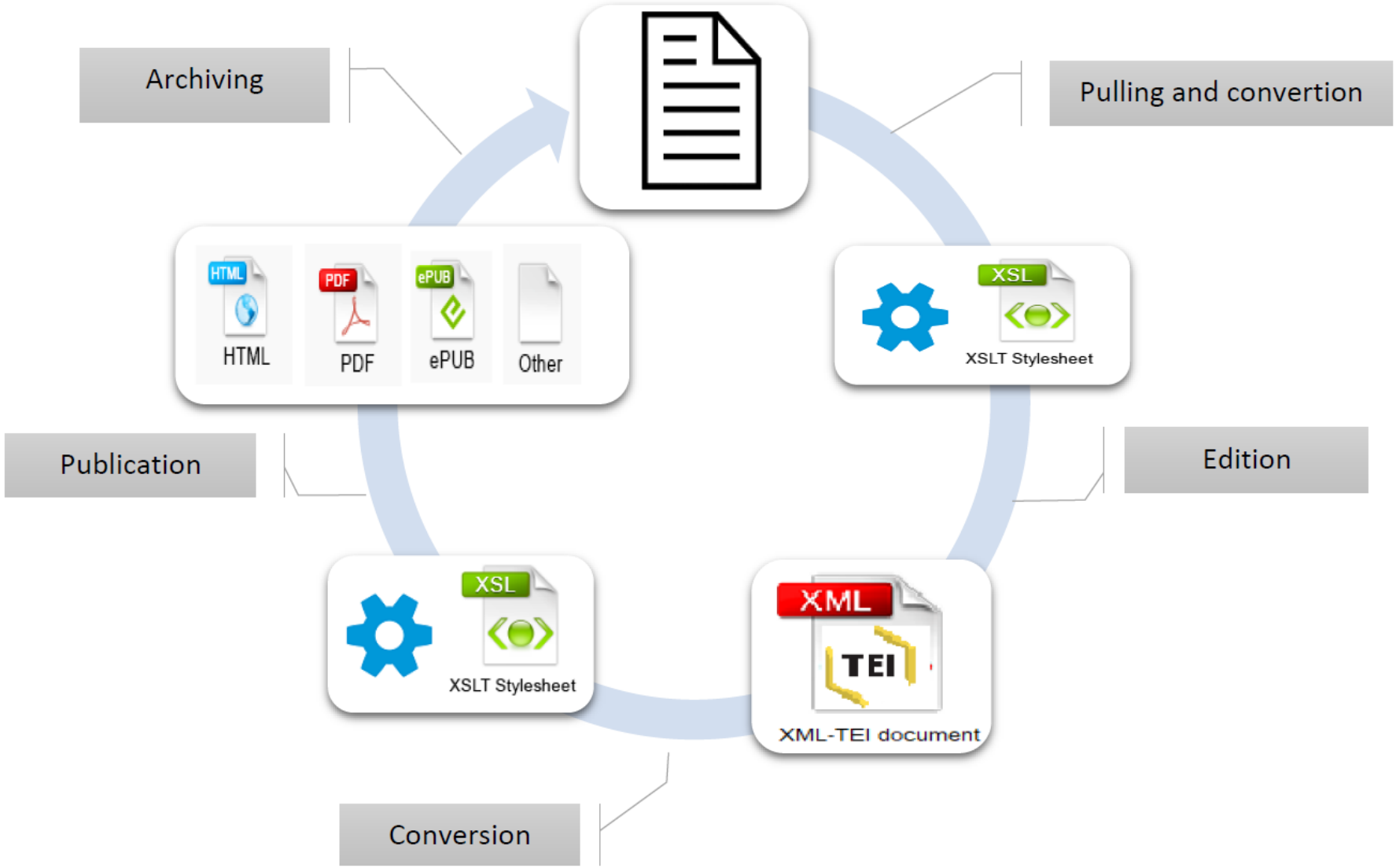
Content management systems

There is also Content management systems for publishing/querying TEI, e.g.

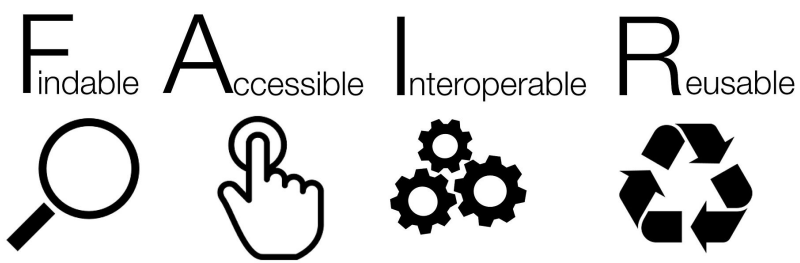
- TEI boilerplate;
- Versioning machine;
- Nemo;
- TXM (desktop or server).

Plan

Lifecycle of the data



FAIR data



(Img. SangyaPundir CC BY-SA 4.0)

- Findable: access on the website of the project; well-calibrated for archiving (Zenodo, Nakala);
- Accessible: open format, without licence;
- Interoperable: open format (again!); standard metadata; consistent vocabulary and conversion stylesheets.
- Reusable: open for re-use and free access (open science!).

Summary

- TEI has been conceived for the need of research projects;
- its manuscript description module was originally made for medieval western manuscripts;
- yet it is highly customisable and can be adapted to a specific project.
- It satisfies the needs and ethics of DH and Open Science;
- but it has a technical "barrier to entry" and also demands a conceptual effort to adapt it.

TEI, a consortium…

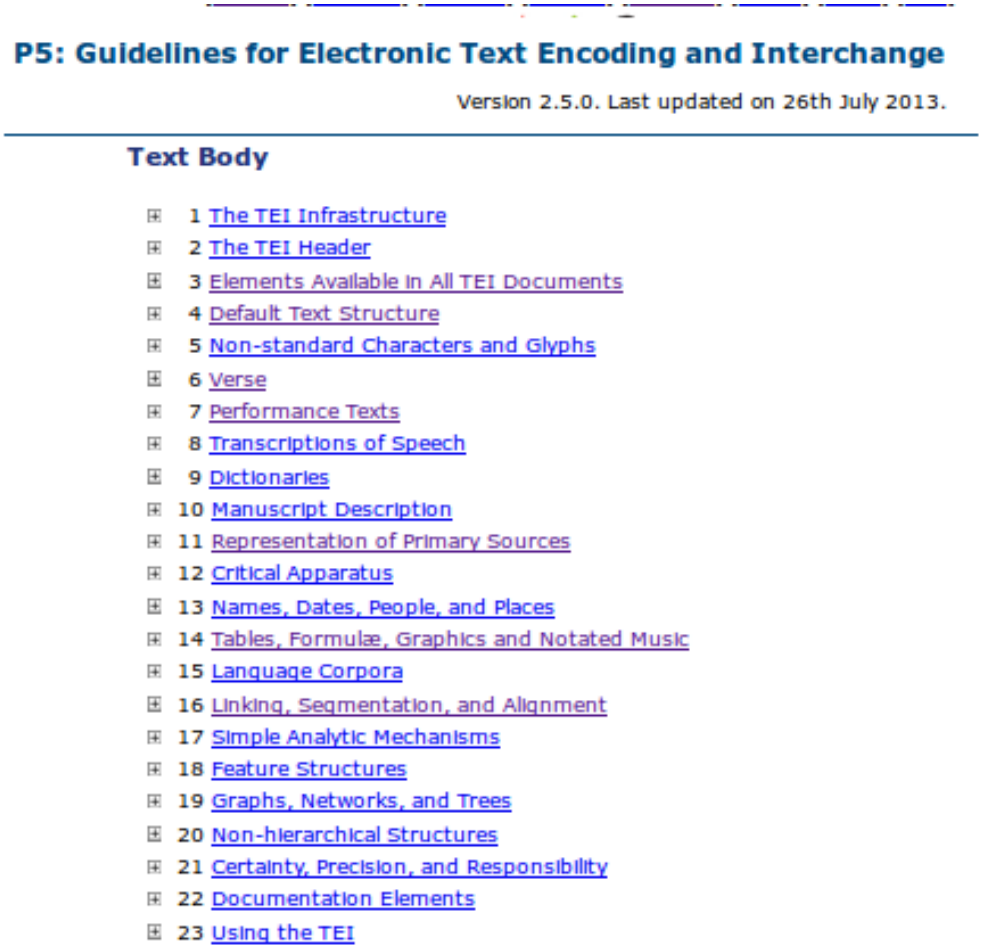
managed by its community

- Board of Directors;
- Technical Council;
- Institutional (e.g. Oxford University…) and individual members;
- Workgroups;
- Special Interest Groups, e.g.
  - Manuscripts,
  - Tools,
  - …
- a *vast* community of users and projects.

and lively

- Annual member meeting (2018: Japan!);
- very responsive mailing-list (`TEI-L@listserv.brown.edu`).

Guidelines



P5: Guidelines for Electronic Text Encoding and Interchange

Version 2.5.0. Last updated on 26th July 2013.

Text Body

The Guidelines

- 23 chapters, on various subjects (including manuscript description);
- technical documentation of elements, attributes, etc.
- plenty of examples, use-cases, explanations...

Projects

Currently 185 projects listed on the TEI website (and probably many times more not listed)

Text editions

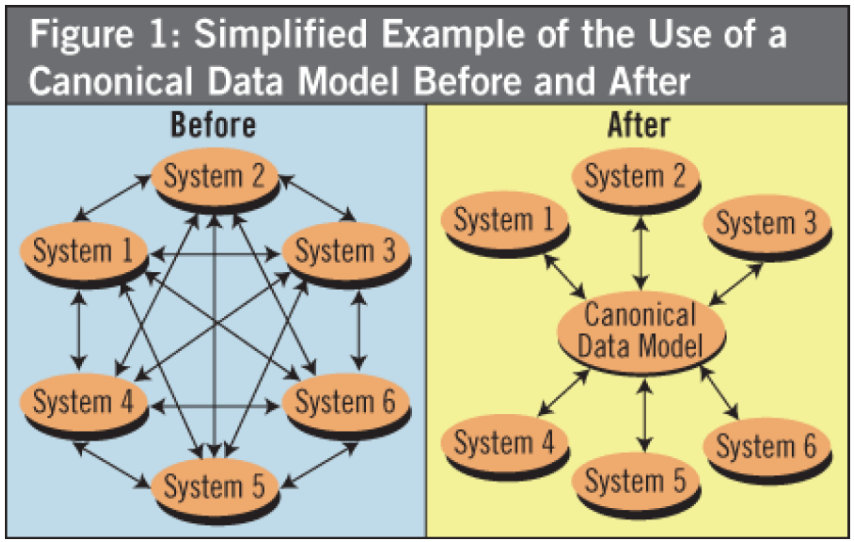
- *Vincent van Gogh - The Letters* (`http://vangoghletters.org/vg/`)
- DicAT: Dictionary of Traditional Agriculture: English-French-Chinese-Japanese (`http://dicat.huma-num.fr/dicat/presentation`).

Manuscript catalogues

- **E-Codices**: Virtual Manuscript Library of Switzerland (`http://www.e-codices.unifr.ch/en`);
- **Manuscriptorium** (`http://v2.manuscriptorium.com`);

An XML implementation and tools

Canonical data model



(from Steve Hoberman, « Canonical Data Model », Information Management Magazine)

A variety of tools

- transformation stylesheets from / to a variety of formats (xhtml, LaTeX, ePub, ...);
- corpus management tools;
- publication tools and CMS (e.g. Lodel);
- analysis tools (e.g., TXM).