Modeling and transforming a multilingual technical lexicon for conservation-restoration using XML

Alice Lonati[1], Violetta Lonati[2], and Massimo Santini[2]

[1] Associazione Giovanni Secco Suardo – Lurano, Italy
resi@associazionegiovanniseccosuardo.it

[2] Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano – Italy
{lonati,santini}@dsi.unimi.it

LMCR project The European Project *Lessico tecnico Multilingue di Conservazione e Restauro*, started in 2001 and still active, is aimed to eliminate the confusion on terms used in the conservation-restoration field, by establishing a scientifically sound lexicon in five languages.

The project carries out historical and scientifical research, resulting in completely original texts. Hence, it does not belong to the field of digitalization of existing sources. Its main result will be the realization of a hierarchically organized lexicon of technical terms, also available in digital form.

The lexicon Terms are organized according to a hierarchical structure and described by a record per language comprising data on: ethymology, definition, description, historical and geographical context, facet, synonims, near terms, and equivalent terms in the other languages. Furthermore, illustrations, bibliography references, and authorship information are recorded.

One of the main issues is the fact that any technical term has a rich meaning that may vary strongly in the different languages. Thus, each term is the result of continuous discussions and exchanges among experts, and each term record is not a simple literal translation, but it is a localized version that may present relevant differences in the various languages.

Project organization The project, leaded by *Istituto Centrale per il Restauro* (Roma, Italy), involves as partners some of the main organizations acting in the field of conservation-restoration in Europe. Each partner coordinates a working group formed by experts with humanistic expertise, or technic-scientific know-how in the area of conservation-restoration. Each group is producing term records for a subset of terms in its own language, and localizing term records edited by other groups. The project coordinator *Associazione Giovanni Secco Suardo* (Bergamo, Italy) is charged with gathering, managing, and validating all term records produced by the partners.

Fostering the interaction among experts –necessary to elaborate the lexicon, establish the term properties and their relationship, write and revise them– is an intrinsecally complex task. For this reason, from the start of the project, all partners have agreed that suitable digital tools would have been necessary to support the process.

Our contribution After the set of terms, their features and the hierarchy among them was established by the partners, a precise logical format for the term records was defined. This allowed to base the logistic support of the project on a *center-periphery* schema.

At the *periphery*, the experts contribute their texts strictly respecting the logical format for the content, but without constraints about the physical support (handwritten texts, plain or formatted text files, other).

The *center*, coordinating the work, first receives the texts, then inserts and marks them up, reproducing the hierarchical order. Such tasks can be performed by a person with basic computer skills, by means of a

WYSIWYG tool (see Fig. 1). The same tool can be used also to revise and edit previously inserted data, and to validate automatically the structure of each term and other data involved.

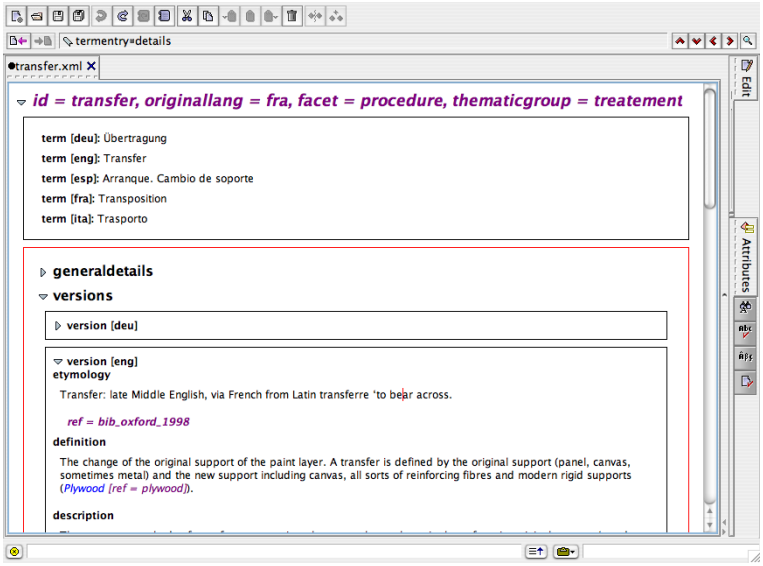

Figure1. Screenshot of the XMLmind XML editor equipped with our CSS stylesheet.

Moreover, we equip the *center* with a set of transformations ready to execute. They produce specific views of the content (for internal use) or put the whole lexicon in a format easy to access and to read. In particular, two different outputs ready to publish can be obtained: the first one consists of a unique file ready to print; the second one is a website (see Fig. 2). Both outputs share the following features.

- They require common non-proprietary software to be read; moreover, they can be easily put on the Internet, on CD, on files to download or copy and look up on one's own computer.
- All pages have a homogeneous visual aspect that can be customized with an independent stylesheet, without re-generating the content of pages.
- The pages are connected in many ways by means of hyperlinks in the web version and cross-references (with related page numbers) in the version to print.
- There are pages containing the term records in the different languages, and pages that give the possibility to access the lexicon according to different approaches (aphabetical indices; table of the linguistic correspondence among terms; inverted indices of terms, authors, bibliographic references).

Notice that, since executing the trasformations is costless, the *center* can produce several drafts of the lexicon during the editing and revision process, and periodically update the published edition with no further production costs.

Implementation details The tools we developed are all based on the eXtensible Markup Language (XML) [1,2]. XML is a descriptive markup language that allows the users to define their own tags. Unlike HTML where tags are used mainly to visually format texts, in XML tags are used mainly semantically, that is according to the content of the text. XML is widely used, especially in the Internet, to share

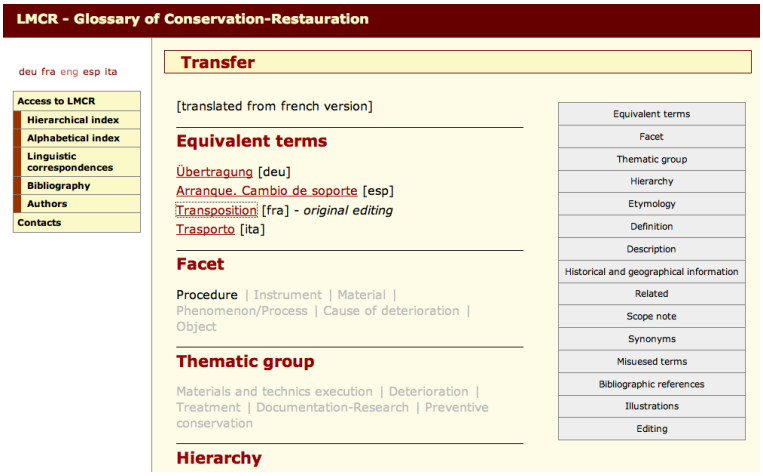

Figure2. A term page of the website produced using our tools.

structured documents across different information systems. For instance, XML is the technology the Text Encoding Initiative (TEI) is based on [3,4].

All data included in the lexicon (term records, bibliography, illustrations and authorship) adhere to a formal strcuture, defined by a specific Document Type Definition (DTD) [2]. To allow the use of formatted text (e.g. paragraph, emphasized text, numbered lists and so on) within some elements, a small set of basic HTML tags have also been included in the DTD. Moreover, the part of the DTD corresponding to bibliographic data is quite detailed, in order to guarantee the possibility to integrate the lexicon bibliography with other bibliographic resources, if needed. A semplified version of the DTD is showed in Fig. 3.

The whole lexicon is included in a unique XML document, containing all term properties. In order to facilitate the editing of a single term, each term record is contained in a separate file, and all files are merged by XInclude directives [5] respecting the hierarchy. Insertion and editing can be made by using a WYSIWYG editor, for instance we suggest the freeware software XMLmind XML editor [6], equipped with a CSS [7] stylesheet written according to the DTD (see Fig. 1 for a screenshot). Notice that this is not a costraining choice, since such operations can obviously be done directly with a generic text editor.

We manipulate the XML content by using the eXtensible Stylesheet Language (XSL), which comprises three languages: XML Path Language (XPath) [8] is used to address the part of an XML document; XSL Transformations (XSLT) [9] is used to describe how to transform XML documents by selecting, ordering, and grouping parts of data; XSL Formatting Objects (XSL-FO) [10] is used to specifying the visual formatting of an XML document. The actual transformations are performed by some XSLT processor, as Saxon [11]. In particular, we designed and implemented transformations to obtain

- some temporary tools useful during the process:
  - to identify the still missing parts of data,
  - to list and order already inserted item according to various key;
- list and tables, like:
  - the table of linguistic correspondence among terms,
  - the alphabetical and hierarchical lists of terms,
  - the lists of authors, bibliographic references, illustrations;

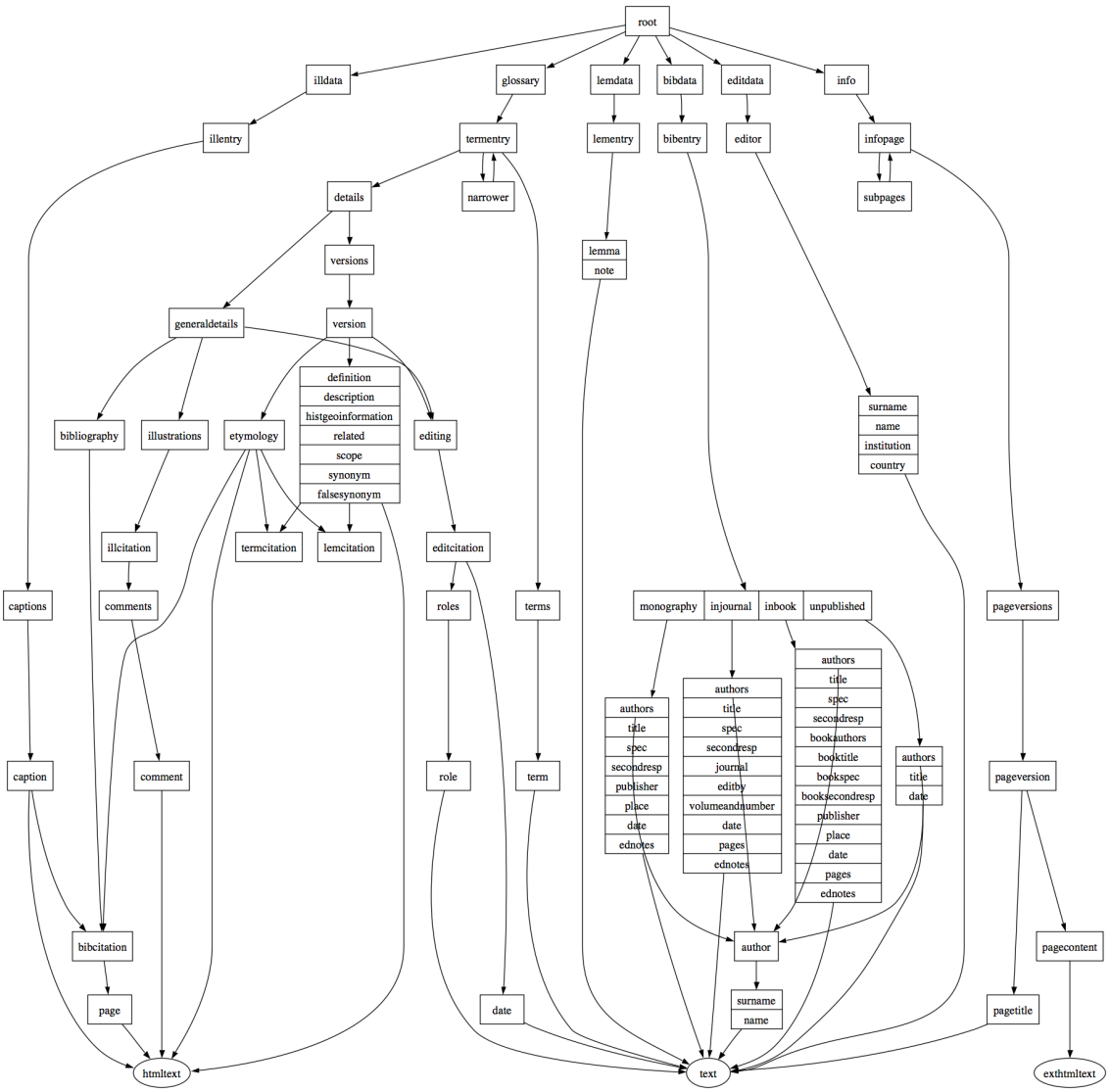

Figure3. Graph representing a simplified version of the DTD.

- inverted indices, like:
  - the index of terms: for each term T, the index provides the list of all other terms whose record cites T,
  - the index of bibliographic citations,
  - the index of authors,
- the whole output to be published:
  - as a website formed by HTML pages,
  - in printed form, from a PDF file.

Final remarks We provided the LMCR project with a set of XML-based tools to model, gather, manage, edit, revise, publish, and update all information concerning the technical multilingual lexicon on conservation-restoration the project is producing. All tools have been designed and implemented in order to support all pahses of the project in a homogeneous way, so that partial results can be directly used in the subsequent phases, without any further effort.

The important issue of portability was taken into account [12,13]. All tools are developed using mainly free software and they fully respect international standards. As a consequence, the lexicon is a a digital resource that can be accessed and used without any special hardware or software restrictions, easily connected or interfaced with other similar sources, exported or transformed in other formats, durable in time.

Finally, our solution has interesting advantages with respect to the users it is target to. The researchers involved in the project have been chosen for their high professionality and long experience in conservation-restoration, but, on the other hand, their acquaintance with computer use is not uniform, if not completely absent. The technical solution proposed takes this context into account, without forcing any change in the usual working methods of the project members, and requiring only a simple training for a pair of operators. However, we noticed a relevant unforeseen effect: the possibility to have easily and quickly several drafts of the lexicon facilitates the exchange among experts, especially in the revision phase, so that researchers are having direct advantages and the quality of their work is improving thanks to this digital support.

References

1. Tim Bray, Jean Paoli, C.M. Sperberg-McQueen, Eve Maler, and François Yergeau. Extensible Markup Language (XML) Version 1.0 (Fourth Edition). http://www.w3.org/TR/REC-xml/.
2. Charles F. Goldfarb. *The SGML Handbook*. Oxford University Press, 1990.
3. Text encoding initiative. http://www.tei-c.org.
4. Nancy M. Ide and ed Jean Veronis. *Text Encoding Initiative: background and context*. Kluwer Academic Publisher, 1995.
5. Jonathan Marsh, David Orchard, and Daniel Veillard. XML Inclusions (XInclude) Version 1.0 (Second Edition). http://www.w3.org/TR/xinclude/, 2006.
6. XMLmind. http://www.xmlmind.com/xmleditor/.
7. Bert Bos, Tantek Çelik, Ian Hickson, and Håkon Wium Lie. Cascading Style Sheets Level 2 Revision 1 (CSS 2.1) Specification. http://www.w3.org/TR/CSS21/, 2007.
8. James Clark and Steve DeRose. XML Path Language (XPath). http://www.w3.org/TR/xpath, 1999.
9. James Clark. XSL Transformations (XSLT). http://www.w3.org/TR/xslt, 1999.
10. Anders Berglund. Extensible Stylesheet Language (XSL) Version 1.1. http://www.w3.org/TR/xsl/.
11. Saxon. http://saxon.sourceforge.net/.
12. Steven Bird and Gary F. Simons. Seven dimensions of portability for language documentation and description. *Language*, 79(3):557–582, 2003.
13. Alan S. Morrison, Michael Popham, and Karen Wikander. *Creating and documenting electronic texts: a guide to good practice*. AHDS Guides to Good Practice. Oxbow Books, 1999.