LISA

Terminology Management

A comparative study of costs, data categories, tools, and organizational structure

LISA™

http://www.lisa.org

Terminology Management A comparative study of costs, data categories, tools, and organizational structure is a publication of the LISA Terminology Special Interest Group, and was written by:

Kara Warburton, IBM Corporation
Paolo Vanni, J.D. Edwards & Company
Inna Geller, Medtronic, Inc.
Raphael Prono, Xerox Corporation

The authors would like to thank Arle Lommel (SMP) for publication assistance.

Terminology Management

A comparative study of costs, data categories, tools, and organizational structure

Contents

Introduction

This document summarizes the results of a comparative study about terminology management processes conducted by the LISA Terminology Special Interest Group (www.lisa.org/term). The study compared costs, data categories, tools, and organizational structure related to the terminology activities within the Special Interest Group (SIG) member companies.

The following companies participated in this study:

- IBM Corporation
- Xerox Corporation
- Medtronic, Inc.
- Cisco Systems Inc.
- J.D.Edwards & Company
- SAP
- Oracle Corporation

The LISA Terminology SIG undertook the current study as a follow-on activity to the LISA Terminology Survey, which was conducted in October 2001. That survey contained more than 70 detailed questions and more than 100 companies responded. Members of the LISA Terminology SIG, all of whom actively manage terminology within their company in a systematic manner, identified the following seven theme topics from that survey for further study. It was felt that these topics would provide the basis for best practices in terminology management.

- Organizational structure
- Costs
- Tools
- Data categories
- Business processes
- Terminology workflow
- Quality objectives and measurement

The SIG began by conducting a comparative study of the practices currently in place in their respective companies. The first four topics were selected for the first stage.

Methodology

A set of questions was prepared for each theme topic. Questions were chosen that would enhance, not duplicate, the original survey. The questions were distributed to the SIG members in spreadsheet format. After all the data was collected, a cross-company comparison was made. The SIG held conference calls to discuss the process and data; minutes are recorded on the LISA Terminology SIG bulletin board[1]. Significant trends and other observations are provided in this report. Also, the process and preliminary findings were presented at the LISA Forum in Heidelberg, Germany, in November 2001.

[1] http://www.lisa.org/sigs/phpBB/viewforum.php?forum=1&2

1. Organizational Structure

Six companies responded to the questions about organizational structure. The six companies varied significantly in size and revenues, from 50 million (USD) to 86 billion (USD), but the following similarities were noted:

- They sold their products and services globally.
- Information technology is a primary business for five of the six.
- Three of the companies are also manufacturing companies.

Terminology management staffing

Three companies have their terminology function staffed with full-time employees at the corporate level, thus recognizing this function as an important component in the total quality process. Two of the companies employ terminologists on the divisional level and the remaining company employs part-time terminologists for various projects. None of the companies outsource their terminology management. Employees managing the terminology come from various backgrounds, including linguistics, localization, translation, and technical writing. All are active members of the project management teams, while none directly manage staff.

Degree of process integration

Four of the six companies have an internal stand-alone terminology database. Two of the companies have their terminology data integrated within their localization systems, two have it integrated with a product classification system, and one has it integrated with an information retrieval system.

2. Costs

Seven companies responded to the questions about costs.

Cost of terminology staff

Terminology teams are generally small. The average team size is three people, with one particular exception of a team of seven.

Terminology staff is also expected to manage projects. The average proportion of time spent on project management is 29 percent, and the maximum proportion is 50 percent.

On average, the terminologist's time is distributed as follows:

- Project management: 30 percent
- Terminology extraction: 22 percent
- Review of terminological entries: 16 percent
- Creation of new terminological entries: 18 percent
- Creation of derivative products (glossaries, Web sites, reports): 14 percent

Individual responses vary significantly from the averages. Nevertheless, this figure indicates that terminologists are expected to be flexible and take on project management responsibilities as well as more traditional terminology–related activities. Eighty percent of terminology staff is experienced, showing that terminologists are expected to possess specific knowledge and accurate training in their field and that few terminology tasks can really be delegated to inexperienced staff.

Terminology work also involves non-terminologists:

- Four companies employ a technical expert to solve terminology queries (one paying up to 1600 person-hours/year for that service).
- Four companies employ a technical support specialist for the terminology tools (up to one full-time position).
- Three companies employ a database administrator.

Other roles involved in terminology work include project managers, content creators, editors, technical reviewers, subject matter experts, and marketing personnel. Respondents found it difficult to quantify the level of their involvement.

Translators are always involved in updating the terminology databases: all respondents explained that they expect input from the translation community to contribute to the terminology records (up to 20,000 person-hours/year)

Cost of software licences/development

A majority of respondents (five out of seven) declared that they used internally developed tools for their terminology processes. When they exist in a company, internal tools represent more than 90 percent of overall terminology tool usage. Understandably, the popularity of this type of tool is due to their low or non-existent purchase costs as well as the flexibility they give to users when it comes to development of new features or technical support.

However, commercially available tools are also used by five out of seven respondents. Some companies seem to make a limited use of those tools (ten percent of overall terminology tool usage) while others use them extensively (80 to 90 percent)

These two results suggest that generally companies prefer to use a mix of internally developed and commercially available tools, according to the needs and the facilities provided by the available tool and the cost involved.

Internal infrastructure costs

Almost all terminology repositories are kept on a dedicated server (six of seven) and some respondents spend as much as $5,000 per year on maintenance. External access is not a priority at the moment since only one respondent intends to implement a terminology repository accessible on an external server.

Term collection costs

Manual term extraction is performed by all respondents at varying degrees. Automatic term extraction is less frequently performed. Automatic monolingual extraction is performed by three of the seven respondents and only two of them do multilingual term extraction automatically.

If extraction is in demand but manual processes are preferred to automated ones, it raises the question of the accuracy and the quality of the deliveries of the terminology extraction tools currently available on the market.

Other terminology management tasks (find/process duplicates, find incomplete entries, export data in different formats, research concepts, add definitions, clean up extraction results, translate terms, work with translation memories) are performed by all companies, although at varying degrees.

On a general level, it seems that estimating cost per word is probably not the best way to compare how the various companies operate. Too much data coming from various respondents was inconsistent and could not be compared. This is mainly due to the diverse activities covered by the terminology teams and the scope of work understood under each item. The main conclusion of this section is simply that it needs more study: the Terminology SIG is keen to go deeper into details and make a link between terminology management best practices and what they should cost.

3. Tool

Seven companies answered 45 questions about tools, describing nine tools..

All tools except one are proprietary. This explains why most do not offer a public API. Two companies used the same commercially available tool, but one of those companies extensively customized it to meet its needs. One tool was still in the design phase. Most tools run on Windows® 2000 and Windows® XP, on standard PCs. Most are Web-based and can run over a network. The backend database is of various formats, from proprietary to IBM DB2®, Microsoft SQL Server, and Oracle® Database.

Common features

The features listed below were supported by most of the tools surveyed:

- Text features:
  - Bidirectional scripts
  - Unicode
- User interface features:
  - User-configurable interface
  - Pick lists for data entry
  - Addition of new data categories
  - Sophisticated search capabilities (filtering , fuzzy logic, full-text searching)
- Export features:
  - Export to other formats (but not to TBX)
  - Export of subsets through filters
- Term entry structure:
  - Equivalency of term status (all terms are equally treated in the database with regards position in the entry structure and availability of fields to describe the term)
  - Concept orientation
  - Relationships between terms
- Database administration features:
  - Backup and restore
  - Data replication
  - Record locking

Uncommon features

The features listed below were supported by few or none of the tools surveyed

- Scripting
- Term extraction
- Localizable user interface
- Export to TBX
- Data encryption
- Compliance with ISO 16642
- Concept and lemma view
- Communication functionality, such as e-mail and FTP
- Ontologies

Additional results

- Only four of the nine tools can be integrated with a translation memory system.
- About half of the tools support automatic validations.
- About half of the tools provide workflow functions.
- Some tools support graphics and media attachments.
- Some of the tools support version control.
- Less than half of the tools provide reporting capabilities.
- Some tools provide additional functionality such as tracking, scheduling, and statistical analysis.

It is impossible to draw irrefutable conclusions from this information without conducting one-on-one interviews with the companies surveyed. For example, it is impossible to determine the exact reason why a feature is uncommon. What follows are some possible interpretations of the data.

Observations on common features

Since the companies surveyed all fit the profile of large IT companies that were willing to invest in a sound terminology management strategy, one can conclude that the features selected by most respondents are not simply a coincidence; they are the core features of a robust terminology management system. Most are not surprising, for example, the support of Unicode, sophisticated search, basic database administration functions, and export capabilities. Most also validate the findings of the LISA survey. However, a few warrant additional comments and suggested conclusions.

- Customizability is important: user-defined pick lists, user-configurable interface, and the addition of new data categories.
- Concept orientation and equivalency of term status are two key principles of terminology management promoted by ISO. These features are critical.
- Relationships between terms are also important. This requirement may increase pressure for the tools to provide more advanced terminology relationship management functions, such as for support for multidimensional ontologies.

Observations on uncommon features

The lack of support of TBX could be due to the fact that TBX is a new standard, or it could be because industries do not need to exchange terminology at this time. The same could be said about the lack of compliance with ISO 16642.

The lack of integrated communication features suggests that communication functions such as FTP and e-mail that are already available outside of terminology management tools are meeting basic needs.

Since the LISA survey found that term extraction is a very important terminology activity, the lack of term extraction functions suggests a gap between functionality and need. Addressing this gap will increase the productivity and quality of terminology management activities.

Since it can be said that most companies do not externalize their terminology data, and when they do, they generally do not view the terminology data as confidential, protection features such as encryption are not needed.

Observations on additional results

Few people in the localization industry would disagree that integrating terminology tools with translation memory tools would be beneficial, yet less than half the surveyed tools provide this integration. This is clearly a weakness of the current tools.

Many of the features in this section, which are supported by approximately half of the surveyed tools, refer to more sophisticated aspects of terminology management that respondents to the LISA survey indicated were important but usually missing from the commercial tools: workflow functions, automatic validations of term entries, version control, report generation, tracking, scheduling, and statistical analysis. While such functions are normally absent from off-the-shelf tools, a significant number of these respondents certainly felt that they were worth developing in their own tools.

4. Data Categories

Seven companies responded to the questions about the data categories that they record in their terminology repository. They were asked to select which data categories they use from the list of standard terminology data categories in ISO 12620, and to provide additional information about those data categories. The following is the list of data categories used, followed by the number of respondents who use each, in descending order.

- Term – 7
- Definition – 6
- Source – 6
- Abbreviation – 6
- Full form – 5
- Subset – 5
- Cross-reference – 5
- Usage note – 5
- Input date – 5
- Modification date – 5
- Creator – 5
- Part of speech – 5
- Context – 5
- Note – 4
- Product name – 4
- Gender – 4
- Subject field – 4
- Superordinate concept – 4
- Subordinate concept – 3
- Updater – 4
- Term type – 3
- Synonym – 3
- Acronym – 3
- Variant – 3

One of the respondents did not select many data categories from the list and seemed to be describing the output from a term extraction process rather than a terminology database. Another respondent was describing a database that was in the early stages of design. This explains the lower than expected usage figure for some data categories, such as the part of speech, which is normally considered mandatory in terminology repositories.

The term and definition are the most commonly recorded data categories. All respondents categorize terms into logical groupings through various different data categories (source, subset, product name, and subject field). Five of the seven record variant terms in some capacity (abbreviations, acronyms, and other variants). Taking into consideration the lack of data provided by two respondents, one can assume that variants are widely recorded in terminology databases.

Terminology relations are also recorded by most respondents, in the form of cross- references, superordinate concepts, subordinate concepts, and synonyms.

Textual descriptions in addition to the definition are also recorded by most respondents in the form of context sentences and notes.

Only three of the respondents provided the additional information requested about the selected data categories: markup style, parent data category, data type (free text, pick list, numeric, etc.) field length, and example. This provided insufficient information to draw conclusions about the nature or structure of the terminological entries.

These findings generally confirm the results of the first LISA terminology survey, with one minor exception. In the LISA survey the context sentence was as important as the definition. In the current poll, the definition is slightly more important. This can be explained by the fact that the respondent profile in the LISA survey and the current study are different. The LISA survey included responses from a large number of small localization companies, many of whom would not have the time to create definitions and would find a context sentence an acceptable substitute for translation purposes. In the current study, respondents are large IT companies who have a broader scope of terminology use beyond translation, which would justify the effort to record definitions.

Trademarks

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

IBM and DB2 are trademarks of International Business Machines Corporation in the United States, other countries, or both.

Other company, product and service names may be trademarks or service marks of others.

LISA™

Published by SMP Marketing and the LISA - July 2003

The Localization Industry Standards Association
7 Route du Monastère • CH-1173 Féchy • Switzerland
Tel: +41 (21) 832 32 10 • Fax: +41 (21) 821 32 19 • lisa@lisa.org