

How Debian GNU/Linux is translated into Spanish

Jesus M. Gonzalez-Barahona
Javier Fernández-Sanguino Peña

Resum: La Debian GNU/Linux és un dels paquets de programari per a Linux que compta amb una major distribució. Inclou milers de paquets de programari lliure (free, open source) provinents de moltes fonts diferents. Una de les peculiaritats del projecte Debian rau en el fet que funciona gràcies a un grup de voluntaris seleccionats per ells mateixos. En concret, això vol dir que les tasques de localització (inclosa la traducció) a moltes de les llengües previstes es duen a terme a partir de l'esforç de voluntaris. Aquest article descriu com es localitza el Debian al castellà, com a exemple dels esforços que s'estan fent en el món del programari lliure.

Paraules clau: Programari lliure, localització, anglès al castellà, gestió terminològica, control terminològic.

Resumen: La Debian GNU/Linux es uno de los paquetes de software para Linux que cuenta con una mayor distribución. Incluye miles de paquetes de software libre (free, open source) que provienen de muchas Fuentes diferentes. Una de las peculiaridades del proyecto Debian radical en el hecho de que funciona gracias a un grupo de voluntarios seleccionados por ellos mismos. En concreto, esto quiere decir que las tareas de localización (incluida la traducción) a muchas de las lenguas previstas se llevan a cabo a partir del esfuerzo de voluntarios. Este artículo describe cómo se localiza el Debian al castellano como ejemplo de los esfuerzos que es están realizando en el mundo del software libre.

Palabras clave: Software libre, localización, inglés al castellano, gestión terminológica, control terminológico.

Abstract: Debian GNU/Linux is one of the largest Linux-based software distributions, including thousands of libre (free, open source) software packages coming from many different sources. One of the peculiarities of the Debian project is being run altogether by a self-selected group of volunteers. In particular, this means that all the localization tasks (including translation) to the many supported languages is run on volunteer effort. This paper describes how Debian is localized to the Spanish language, as a case example of translation efforts in the libre software world.

Key words: Libre software – free software – open-source software – localization – english to spanish – terminology management – terminology control.

1. The Debian project

Debian GNU/Linux is one of the largest (if not the largest) Linux-based software distributions. Its main section includes only libre software¹, as mandated by their social contract (a binding document for all the members of the project). Currently, Debian includes more than 10,000 different source software packages (each corresponding roughly to one application), and is used (directly, or by means of derived distributions such as Ubuntu or GNULinEX) by millions of users worldwide.

The Debian distribution is produced by the Debian project, composed by hundreds of Debian developers. All of them contribute as volunteers to the project, usually maintaining one or several software packages. 'Maintaining' means following the actual development of the software, and producing Debian packages with each new release. The actual development happens in the "upstream" project (that is, the original project which actually produces and maintains the program). The job of package maintainers, therefore, consists mainly in adapting the releases produced by upstream projects so that they comply with Debian specifications about build and installation automated procedures, package descriptions, etc.

In other words, Debian developers obtain libre software programs from the original producers, and organize them together in a coordinated distribution, suitable for end users. An important part of this process is localizing the software for different languages and cultural contexts. In addition, Debian developers also maintain many documentation related to the project (from installation manuals to a complete web site), which also needs to be translated (currently, to several tens of languages).

One of the main goals of Debian is to be transparent for its users. Almost everything about the project is public (except for data with privacy implications, security information with impact on third parties, and some other few exceptions), and publicly available in the Internet. This includes the activities related to translation, which are coordinated in public mailing lists, and tracked using information publicly available from the project web site.

The Debian project uses a formal organization as an umbrella for its legal and financial activities (for example, receiving donations or dealing with copyright issues). This organization is "Software in the Public Interest" (SPI), formed by Debian developers and incorporated in the United States of America.

2. The Spanish language localization team

In Debian, each supported language has an associated localization team. For Spanish, it is the *debian-I10n-spanish* team, named after the mailing list used for coordination. It is composed of volunteers with varying levels of commitment, ranging from regular contributors with several years of experience to sporadic contributors that come and go. Most of them are not professionally involved in translation issues, neither have studies in the area of linguistics. Some of them are Debian developers, while others are persons interested in libre software in general, or in Debian in particular. Usually they are trained in IT-related issues, and in many cases they are capable of writing software for automating translation-related issues.

The main tasks performed by the *debian-I10n-spanish* team are:

- Translation of documents of the Debian Documentation Project (DDP, documenting different aspects of the Debian distribution or the Debian project).

¹In this paper "libre software" will be used to refer both to "free software", as defined by the Free Software Foundation, and "open source software", as defined by the Open Source Initiative. In summary, this means that programs can be freely used, studied, redistributed and modified by those obtaining them.

- Translation of the Debian web pages.
- Localization of packages (usually applications or modules supporting applications).
- Localization of debconf templates (used for the installation of packages).
- Translation of manual (man) pages, each with information about some program.
- Localization of the Debian installer (the system that installs a Debian distribution in a computer).
- Translation of package descriptions.
- Coordination issues, and decision making in the area of the translation and localization of Debian into Spanish.

For most of these tasks, a mixture of automatic, semi-automatic and completely manual procedures are followed. All of them are carried out by the same community of contributors, but with their own peculiarities, due to the different kinds of information involved. For example, in the case of software packages coordination with the corresponding Debian maintainer and with the upstream package is important, while translation of web pages is only related to the web site maintenance.

The translation team is coordinated by a person, not surprisingly named “coordinator”. His work consists mainly in helping to reach consensus in the team, in detecting problems and areas subject to improvement, and in proposing corrective actions to the rest of contributors. Coordinators are selected by informal meritocracy among regular contributors (usually, the ceasing coordinator proposes the new one). The coordinator also decides which new contributors may have write access to some CVS repositories (such as the web site CVS repository).

3. Sources of texts to translate

Before entering into the details of how the Debian Spanish translation team works, it is worth reviewing briefly the sources of texts to be translated. There are two different situations:

- The text to be translated is produced by the Debian project itself. This is the case of the documents in the DDP, the Debian web pages, the debconf templates, the Debian installer, and descriptions of packages. In all of these cases, the translation is completely produced by the Debian project, and no external coordination is needed.
- The text to be translated comes from another project, in many cases accompanied by translations (although maybe low-quality or not up-to-date ones). The most frequent case is software packages, which are produced by upstream projects.

In this second case, coordination with external parties is needed and encouraged by the Debian project. On one side, translations coming from the original source of software packages, or from other third parties, have to be reviewed to decide about their quality, completeness (when not all the original text is translated) and timeliness. On the other, any change to the translations (or any new translation) should be submitted to the original authors of the package, to avoid divergence between the two versions. This is in fact the recommended procedure, to the point that direct submissions of translations to the Debian maintainer of the package is discouraged: they should be submitted (and accepted) by the upstream project before entering Debian.

Because of this, tight coordination with the external project is needed, which can cause some problems if, for instance, their translation teams follow different procedures. However, and probably due to the fact that volunteers can work for several projects, this is not usually

the case, and translations (and translation policies) usually flow nicely from one project to another, and back.

4. Tools used

Localization and translation is a part of the software development and maintenance processes. Therefore, translators in libre software projects tend to use some of the tools and systems provided by the project to improve and automate parts of the development and maintenance processes. In the case of Debian, these tools and systems are:

- Bug tracking system, for contributing new localization files, for reporting errors in translations or for contributing fixes.
- Mailing lists, which are the preferred way of communication between people contributing to the translation effort both within a specific team and for coordination between the different language teams..
- Version control repositories, for storing translated documentation, and translated web pages. These repositories (CVS and Subversion are commonly used) permit the storage and retrieval of versions of each file. Translators use repositories for the translations of the web site, for Debian documents and to commit updates directly to localized programs.
- Web infrastructure, which is used in two ways. On one hand it is used to publish the translated documents so the public in general can use them. On the other, it is used to provide interfaces for some translations projects, such as the DDTP², or to provide information useful for coordination, such as the status of translations of different contents (for example, the status of Debconf template translations³).

In addition, there are some specific tools for the use of translators (in many cases, built by translators themselves):

- Coordination robot. Translation of any element (be it a a document or a program) can go through different phases until it is is complete. The robot is a way of keeping track of the status of those elements to facilitate coordination and prevent duplicate work. The coordination robot works in combination with the mailing list, monitoring its messages. When the subject line conforms to some specific patterns, the robot interprets it as a command, and acts according to its instructions. The command can inform of the intent to translate or review, request for a review, inform of chances for commenting, etc. With this information, the robot produces some web pages with information about the status of all the texts in process of translation⁴. It also reports to the bug tracking system (and checks for further updates to the report) when needed.
- Scripts for automating translation-related tasks, such as checking whether translations are up to date, detecting packages without localization or with old localization files, etc.

²<http://ddtp.debian.net/ddtss/index.cgi/es>, last visited on June 30th, 2008

³<http://www.debian.org/international/l10n/po-debconf/es>, last visited on June 30th, 2008

⁴<http://www.debian-es.org/cgi-bin/l10n.cgi?team=es>, last visited on June 28th, 2008

- Scripts for producing information available via web, such as: summary information about the translation of the web site to all supported languages⁵, specific information about the status of the translation into Spanish of all web pages⁶, statistics about the status of the translations of the Debian installer⁷, status of translations of software packages⁸, etc.
- Tools to automatically generate translation compendiums for the language, to speed up the translation process by using previous work. These compendiums⁹ can be used by programs supporting the translation process, so that lines that have already been translated elsewhere are directly included (either because of a perfect match, or because enough similarity is found).

All these specific tools usually rely on some markup being present in the translated files (such as those for the web site), or in their name, location or meta-information (such as localization files of software packages).

An specific case worth mentioning is the localization of software packages. In Debian, packages normally use the gettext¹⁰ system for internationalization. In this case, all the strings subject translation are marked in the source code so that a specific program (*msginit*) can extract a file ready to be translated to an specific language (the *.po* file for that language). The work of the translator is therefore to go over the *.po* file and translate any missing string. The user will later select, at run time, the language to use, by setting an environment variable (nowadays, usually in the "Preferences" panel).

The libre software community has produced many different tools to handle gettext *.po* files, including GUI tools such as *ktranslator* and *gtranslator*, which greatly assist translators since they can forget about the specific formats used. In addition, these tools are prepared to work with translation compendiums (as described above) to speed up the process by using already available translations.

In Debian the gettext system is used not only for the translation of upstream software programs. It has also been extended and adapted for its use in other areas. The debconf templates used to configure packages when they are installed, or many documentation (both manual pages and documents produced by the DDP) also use gettext, through the use of specific scripts that convert markup documents into *.po* files that translators can better handle.

5. Main processes

To become a part of the team, it is usually enough to contact the coordinator, and start contributing. Once a person is a part of the team, the process used for the translation (or updating of a translation) of some text may vary, but usually follows the following structure:

- A contributor decides to translate some document. For informing the rest of the team, an "Intention to Translate" message is sent to the mailing list. For selecting the text to

⁵<http://www.debian.org/devel/website/stats/>, last visited on June 28th, 2008

⁶<http://www.debian.org/devel/website/stats/es.html>, last visited on June 28th, 2008

⁷<http://d-i.alioth.debian.org/l10n-stats/>, last visited on June 28th, 2008

⁸<http://www.debian.org/intl/l10n/po/es>, last visited on June 28th, 2008

⁹Such as the ones available at <http://i18n.debian.net/~eddyyp/compendiums/po/es/>, last visited on June 28th, 2008

¹⁰<http://en.wikipedia.org/wiki/Gettext>, last visited on June 28th, 2008

translate, the contributor may consult the different web pages with summaries and statistics about the status of translations and translations missing or not up-to-date.

- When the document is translated (or the translation updated) it is sent to mailing list attached to a "Request for review" message.
- Upon reception of a request for review, eventually some other contributor reviews it, and sends it back to the list, as a reviewed document.
- At this point, the document is ready for submission. If it is a web page, the coordinator will upload it to the CVS of the web server, which means that it will be automatically uploaded to the web site. If it is a localization for a software package, it is submitted to the upstream project, which hopefully will accept it. In this case, the translation will enter Debian when the next release produced by the upstream project is packaged by the Debian maintainer. Only if the upstream project is Debian itself, a bug against the corresponding package is reported, with the translation attached. Eventually the Debian developer in charge of the package will address and fix it, by including the translation in a new release of the package.

All the messages sent to the mailing list follow the conventions needed to be interpreted by the coordination robot, so that it can track the status of the translation of each document.

In the case of the localization of packages, the process includes contacting the upstream project to contribute the translation. This can be done either by the debian-i10n-spanish contributor, or by the Debian developer maintaining the package.

Another process which is specially interesting is how errors in translations are reported and fixed:

- Some person (for example, a Debian user) notices an error in a translation.
- That person reports a bug (using the bug tracking system) to the corresponding Debian software package. That bug report should be tagged as "i10n" (a specific tag for translation-related issues).
- If the subject of the bug report also include "INTL:es", some of the tools to gather the state of translations and proposed translations will easily detect and track it.

Once a bug report is in the bug tracking system, it can be easily monitored, and will stay documented and public.

6. Translation policies and terminology management

One of the challenges faced by the Debian translation team is to agree both on the terminology and on the translation choices to use, specially having into account that all participants are volunteers (which means that enforcement of rules cannot be ensured by contract) and culturally and geographically diverse. However, the translations produced are quite consistent, which means that both formal and informal means work well enough even in this environment.

The most prominent formal resource helping in this coordination is the section "Normas correspondientes a las traducciones" of [1], which can be considered as a brief style manual. It describes the main rules and recommendations that have been agreed by the group (usually, by discussion and consensus agreement in the mailing list). It includes:

- A short set of general recommendations, that could be considered as a brief manual of good practices in Spanish (for the project). Those recommendations range from "use the formal 'usted' instead of the informal 'tú' " to "avoid false friends" (providing a list of common cases).

- A set of resources (talks, articles) explaining the Debian translation project itself, which should be known to contributors.
- Some general resources about the Spanish language, such as the “Diccionario de la Real Academia de la Lengua Española”. In this respect, it is interesting to note that normative documents coming from the Real Academia de la Lengua Española (RAE) are considered (not only here, but also in discussions in the mailing list) as mandatory for the project. No reference to other well known dictionaries (such as Maria Moliner) was found in the formal reference documents, and the RAE opinion is usually preferred to any other.
- Some specific texts about translation and technical translation.
- A “Debian glossary”, with terms specific to the Debian project. Agreement about these was reached in the mailing list, usually with little trouble, and early in the life of the translation team.
- A list of glossaries, complementing the previous one. Those are maintained either by specific collaborative projects (such as ORCA¹¹, which is mandatory for IT terms), or by libre software projects (such as those of the GNOME or KDE projects).
- A list of style manuals, which are expected to complement the project style manual itself. Those are also collaboratively maintained by groups related to libre software.

Not surprisingly for a project like this, all the references and resources considered are available online for free. This not only permits a comfortable and quick use, but also have into account that many contributors, not being professional translators, probably do not have access to specialized books except for those freely available in the Internet.

Several procedures are used to enforce these rules. The first one is their voluntary observance by contributors. Since the rules were previously agreed by consensus, this is encouraged by the group, which can track errors and problems by submitting specific messages to the mailing list (being this another enforcement procedure). Those can be later addressed by the original translator, or by other contributor.

Another point of enforcement is the inclusion of the translation in the official CVS repository of the translation team (when this applies). Only a small group of persons can write new content in that repository, and therefore all translations are in fact taken from the mailing list by them. Before writing the to CVS, they can check the text, and at that point decide that some rule has not been followed.

However, translations (localizations) of programs are not included in the CVS, but sent to the Debian developers maintaining the corresponding packages. They usually do not speak Spanish, and therefore cannot check the translations. Therefore, they tend to rely on

contributors, and are often confident that any error will be detected later (and soon) by other Spanish speakers using the program.

From time to time, discussions about specific terms arise. If those terms can be found in one of the referenced glossaries, or in the Dictionary of the RAE, the issue is usually settled that way. Otherwise, a discussion starts in the mailing lists, hopefully reaching to an end when the original poster adopts a decision, and the rest of the list agrees by not opposing. In some cases the coordinator can decide to include this new term in the glossary maintained by the group if it is clearly specific to Debian. Otherwise, mail messages with the discussion can be

¹¹<http://quark.fe.up.pt/orca/>, last visited on June 27th, 2008

referred later, if the same issue arises once more. Therefore, in some sense, the mailing list is acting as a glossary.

7. Credits to translators

Since most of the management of translated documents is done with automatic tools, it is not difficult to maintain information about who translated what. When the coordination robot is used, for example, it is easy to track who (and when) produced the original translation, who reviewed or commented it, when the translation was modified, etc. Therefore, authorship of translations can be traced with great detail.

However, in most of the cases, that information is not included in the translated texts themselves. This causes that persons reading or using them cannot easily know who was the translator. In fact, it is difficult to notice, for a casual reader or user, whether a certain translation or localization was produced by somebody collaborating with the Debian project, or elsewhere (maybe by the upstream developer team). This situation is consistent with that of libre software in general, where the casual user cannot, in many cases, know who developed a certain program (even when usually that information can be retrieved with fine-grain details from the project information platform).

One of the most relevant exceptions, when detailed information about the translators is provided to readers of a document, is the Debian Documentation Project. In that case, translators can (and usually do) mark the documents they translate with their name and contact information, in fact at the same level than the original authors.

Usually, the information provided by the Debian project is copyrighted by Software for the Public Interest, which provides it under a libre license (e.g., the Open Publication License for the website, or the libre software license used by the upstream project for localizations of software packages). Transferring copyright is, however, not an easy matter, subject to many local regulations, and requiring translators to sign waivers. Since translation to the Spanish language is done in many different countries (mainly in Latin America and Spain, but with contributors also coming from Australia, the United States and other countries) it is common for translators to retain copyrights for their work, publishing it under the same libre license used by the original authors so that it can be freely distributed and used.

8. Conclusions

The translation of the Debian system into Spanish is a huge effort. At the time of finishing this paper, more than 185,000 strings present in software packages have been translated¹² (being Spanish the third language only to French, which is close to 225,000 strings, and German, which is slightly over 200,000). More than 2,411 files¹³ (amounting to about 6,225,000

characters) have been translated in the website. And many Debian documents, man pages, etc. are also available in Spanish.

This effort is being accomplished by a large community of translators. Some of them collaborate under the Debian umbrella, while some others contribute directly to upstream projects. And all of them are coordinated with more or less detail, in the end sharing the same translation resources (tools, compendiums, glossaries, etc.).

¹²<http://www.debian.org/intl/l10n/po/rank>, last visited on June 28th, 2008

¹³<http://www.debian.org/devel/website/stats/es.html>, last visited on June 28th, 2008

For the coordination of the needed activities, some formal policies, and the support of some tools (either specific for translators, or more generic for all software developers), are used. The preferred decision mechanism is consensus, but some ways of enforcing decisions are also used. In any case, the fact that the vast majority of contributors are volunteers makes it difficult to work in any other way.

Understanding how this complex mixture of contributors, most of them not professionally experts in translating, can produce a localized system of this size, with a (at least) reasonable quality, deserves a deeper analysis. In this paper, the author tried only to show the main characteristics of the translation process, without entering into the details of why it works, and how it could be improved. This would be subject to further research.

References

- [1] Coordinación de traducción de documentos de Debian al castellano. Technical report, Debian Project, 2008. Last visited on June 27th, 2008
<http://www.debian.org/international/spanish/>.