What better time than the end of the year to sit back and ponder the things that have happened in the past year and wonder what the next year will bring. (I won't even try to talk about the last and next decade!)

There is a lot to wonder and ponder: What has been particularly notable in the past year in our industry? What were the things that riled us up, enraged us, excited us, or disgusted us?

One thing that comes to mind immediately has to be crowdsourcing. I have written about this a bit in the past and have tried to encourage us not to have the knee-jerk response that so many of us first had when we rejected the encroachment of crowdsourcing into our territory, a territory that we felt entitled to. I've encouraged the translation industry to be actively involved in shaping this (only seemingly) new concept into an opportunity that we can live with and in fact profit from.

Another of last year's notable topics was clearly machine translation. If you have not been involved in several discussions about machine translation with colleagues or other peers this past year, it's time for you to go out and get a social life! Companies like Google, Microsoft, Asia Online, and others have been pushing us to reconsider the applicability of machine translation on the basis of usability. The very concept of quality—which we also have had a love-hate relationship with for a long time—has seriously come under fire. The argument goes like this: Since translation quality is very abstract and arguable (yes, we would all agree here), the only relevant measure for translation is usefulness. For some kinds of texts, high stylistic standards are very important (think: literature, marketing); for others it's accuracy (think: legal, medical); and for still others the only thing that counts is the transfer of information (think: social networks, some technical documentation, customer support data). You may disagree with those classifications, but these are the lines that many
very large corporations are drawing when deciding what to
give to translators and what to have machine translation do.

And then there is another topic that has come charging to the
forefront in just the last few months: the availability of large
amounts of bilingual data that can be used in translation
memories.

Here is just a sampling:

- **MyMemory**: A colossal translation memory of presently
  around 300 million segments that contains data from
  web alignments (app. 30% of the total data), corpora
  such as the EU corpus (app. 50%, see “DGT TM” below),
  and TM contributions from translators. It offers
termology searches, download and upload of
translation memories in TMX (the Translation Memory
eXchange format), editing capabilities for users, and a
strong tie-in to machine translation.

- **BigTM**: A custom translation search engine that can be
  used by LSPs or translators. You can submit the
  translatable text or a sample of it, and the system goes
  out on the web to search for pages similar to the source
  text that already have translations in the target
  language. Within 24 hours it then provides a searchable
  index of the discovered parallel pages that allows you to
  look up how terms or phrases were translated by others
  in the past. (This product is still in its beta phase.)

- **OPUS**: An open-source parallel corpus with a large
  number of bilingual files in many language directions
  containing such varied materials as data from the
  European Medicines Agency, the European constitution,
  the European Parliament Proceedings, the
  OpenOffice.org corpus, the opensubtitles.org corpus, and
  various open-source localization and software
documentation files. The author of the site is a
researcher working in natural language processing and
machine translation, so the files are not especially made
for translation memory—most of them are in a text
format—but they're nothing that could not be converted
to a TM-compatible format or even TMX (and the files for
the European Medicine Agency are in fact in TMX).

- **TAUS Data Association (TDA)**: The TAUS Data
  Association (or TDA) is an association of mostly large
corporate translation buyers who originally came
together to pool their translation memory data to better
train their machine translation engines. TDA has now
just announced that they have launched a relatively
low-priced Professional membership category that allows
you to download 10 times the amount of data that you
upload. Also, as a "by-product" they have opened the
data up to the public as a terminology resource. Both
the terminology searches and the TMX download can be
categorized according to client and a (rather coarse)
subject taxonomy. Presently (December 2009) the
complete corpus includes about 1 billion words.

- **The DGT TM**: The humongous translation memory for
  the Acquis Communautaire (the body of EU law) in 22
  languages and a total of 231 language pairs. It's
  available as a free download and the data is presented
  in TMX format.
Linguee: Linguee is a very large corpus of English-into-German-into-English data (other language pairings will be released in 2010) of web-based translated materials. The web-based data is matched up with the help of a large custom dictionary and other web-based dictionaries. Though the data is not categorized, every entry is accompanied by a link to the originating webpage where webpages or whole websites can be downloaded and aligned (i.e., converted into a translation memory).

And then there are translation environment tools (TEnTs) like Lingotek's suite of tools, Google Translator Toolkit, and Wordfast's VLTM that are built around the concept of anonymous data sharing through translation memories or alignment tools like AlignFactory and NoBabel's AutoAligner that have finally made alignment of large amounts of web-based contents feasible.

So what are we going to do with all of this? Is this sudden flood of data going to be helpful or harmful to our productivity via translation memory technology? The short answer is: I don't know. But I do have an inkling.

When I first started to use translation environment tools (TEnTs), I was very eager to build up my own data so that I could benefit from my past labor. My "Big Mama TM" grew and grew, and I was always excited to find matches from (almost) forgotten previous projects. As the years passed, I continued to use and feed my meanwhile obscenely large Big Mama TM, but her usefulness seemed to decline rather than improve. Too much time had passed between the earlier projects and the current ones to really classify them when matches were displayed (despite every translation unit being described with subject and client information). In addition, language had changed and my skill levels had, too, causing a lot of time to be spent deleting or wading through useless suggestions from the TM. The fact that many of the newer TEnTs now also offered subsegment matching that allowed them to dig even deeper into the language materials did not help either.

I have increasingly come to realize that while large amounts of data are very powerful, they can also be very distracting if they a) originate from a subject matter or client that uses a different terminology or style; b) come from dated or obsolete sources; and c) come from sources with a different quality level.

So what does it mean to have all these gigantic data vaults at our disposal if my conclusions are true? I think that many of them are fantastic as reference materials, but I am just not sure about their value as translation memory data in the classic sense. And it's important to keep in mind that many of these resources were not produced for translation memory purposes (even though that may be their origin), but to feed the ever-hungry statistically based machine translation engines with their favorite food: bilingual data.

Am I suddenly advocating the dismissal of translation memory technology? Not on your life! I still think that TM technology in concert with terminology resources should form the
foundation of the tool kit of every translator who works on functional texts. But I have also come to the realization that raw data, including translation memory data, has no value per se. The value of data for the human translator is in its quality and appropriateness.

Here's to a good and successful 2010 and 20-teens!