[*Translating and the Computer 23*: papers from the Aslib conference held on 29 & 30 November 2001]

Productively Testing the Scalability of MT & TM to the Limit! And Beyond?

Jon Wells
MLT (MultiLingual Technology) Department
SAP AG, Walldorf, Germany
E-Mail: Jonathan.Nigel. Wells@sap.com
Tel: +49(0)6227745886
Fax.: +49(0)62277821174

Synopsis

This paper describes a practical example of MT and TM being used to facilitate the initial and subsequent translation of several hundred documents daily in a corporate environment. These documents make up a knowledge database, which customers can use to solve problems, find answers to their questions, install new features and generally gain information and knowledge about the SAP System. The successful implementation of MT/TM and the flexibility and scalability of both the software and its users have saved many millions of Deutschmarks over a period of 8 years. This system has been successfully developed and enhanced due to the active feedback and close cooperation of all parties involved - the developers, SAP and the translation agencies that work with it.

Introduction & Background

SAP is the world's leading provider of collaborative e-business solutions. With 36,000 installations serving 10 million users at 13,500 organizations in 120 countries across the globe, SAP ranks as the world's third-largest independent software provider. SAP has been in the business of e-business for 29 years and began trading publicly in 1988. Founded in 1972 by five former IBM systems engineers, SAP currently employs more than 23,700 people in more than 50 countries. The company is headquartered in Walldorf, Germany.

Almost half of SAP's employees are actively involved in the process of software development. As software is developed, tested and implemented at customer sites, various questions, problems and other issues occur, which must be documented and processed. To this end, SAP has implemented a workflow system in which any SAP user can send queries to SAP for immediate processing, assigning them to various categories, depending on the issue being raised. These reports, officially known as SAP Customer Messages, are then processed by the developers and returned, hopefully with a solution, to the user. The solution, officially referred to as an 'SAP Note' can often, of course, be reused and must also be stored somewhere in the system for future reference. The Notes are also indexed (full text and by keyword) to facilitate future lookup and reference.

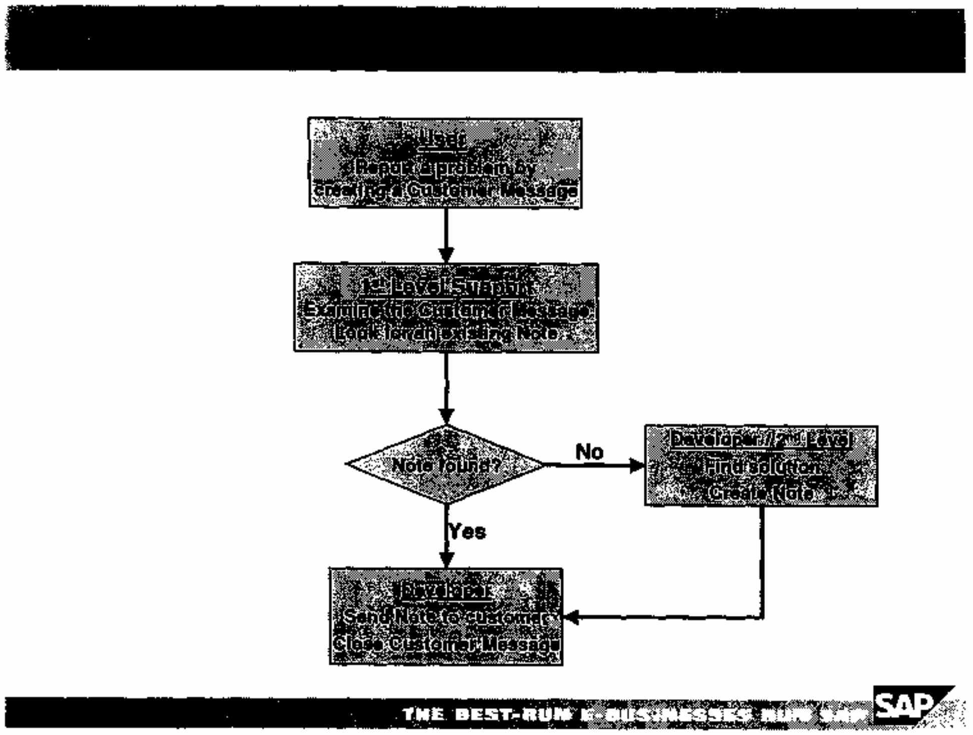


Figure 1: Processing SAP Customer Messages

Figure 1 shows a very simplified representation of the procedure for handling the Customer Messages and the relevance of Notes within this procedure.

Notes can contain a variety of different information and can also be created as a source of information without first receiving a Customer Message. Notes can contain:

- Bug fixes
- Additional documentation
- Process descriptions
- Workaround descriptions
- Recommendations
- Collective information about other Notes
- Other general information

The structure of a Note is as follows:

- Short Text
  One line description of the content of the Note
- Symptom
  The issue that is addressed in the Note
- Keywords
  Keywords used for quick searching
- Cause and Preconditions
  Information required for reproducing the error or information relating to the scope of relevance of the Note
- Solution
  Main text of the Note

In the worldwide market, 60% of SAP's revenues come from the English-speaking world. However, the vast majority of SAP Notes are written by in German by developers based in Walldorf.

This seemingly simple translation problem is severely compounded by a number of factors, which turn it into a fairly major undertaking:

- **The Volume**
  Each Note contains, on average, around 105 words, though they can be as long as 100 pages. Currently, around 200 new Notes are created every day and most of these are released for translation. This gives a total of 21,000 new words to be translated every day, and this volume is steadily increasing - it is approximately 25% higher than the same period 12 months ago. The system contains around 450,000 Notes at the moment.
- **Repetitivity**
  Approximately 350 existing Notes are currently modified each day (information updated or added, errors corrected, etc.) and released for re-translation. This gives a total of 36,750 words, part of which has already been translated.
- **Turnaround Speed Required**
  As many of the Notes are urgently required by companies around the world, it has been agreed that Notes should generally be translated within 24 hours of their creation, some, however, within 4 hours. These timeframes are primarily defined by SAP's Support Level Agreements with our customers and simply passed on to the agencies.

In 1993, the MLT (MultiLingual Technology) department at SAP recognized that this area appeared to be an excellent candidate for the combination of MT and TM and, in conjunction with Siemens, began to implement the **METAL** system to assist in the translation of SAP Notes. Over the next few years, **METAL** progressed and developed, first becoming T1 and then **Comprendium,** moving from the UNIX operating system to Microsoft Windows and from Siemens to SNI, GMS, L&H and then Sail Labs in the process. During this period, the MLT group at SAP provided a complete translation service with full post-editing for Notes, as well as using the MT system for various other large-scale translation jobs within SAP.

In 1996, however, the department began to shift its focus away from translation and at the same time realized that the Notes translation was getting too big to handle in such a small group (5 people at the time). A translation agency (S&D, Rendsburg (now part of Ll0nbridge)) was selected to begin taking over this translation job and, by April 1997, the transfer was complete and the Notes translation was completely outsourced.

In 2000, for various reasons (including Y2K compatibility and a lack of scalability), METAL was officially retired at SAP, along with the Unix-based workflow tools and T1 was implemented. At the same time, a completely new workflow solution (The Notes Translation Solution, or NTS) was developed by SAP to handle the Windows-based procedures. To assist us in this, Sail Labs provided various components to simplify the transition and implementation processes.

As of January 2001, some of the Notes translation has been transferred to a second translation agency (SimulTrans, Dublin).

We estimate that the implementation of an MT/TM-based solution in this case allows us to save 40-60% of the translation costs (depending on how the figures are calculated). Regardless of whether you look at the best case or worse case figures, the saving over the operating period has been in the region of several million Deutschmarks.

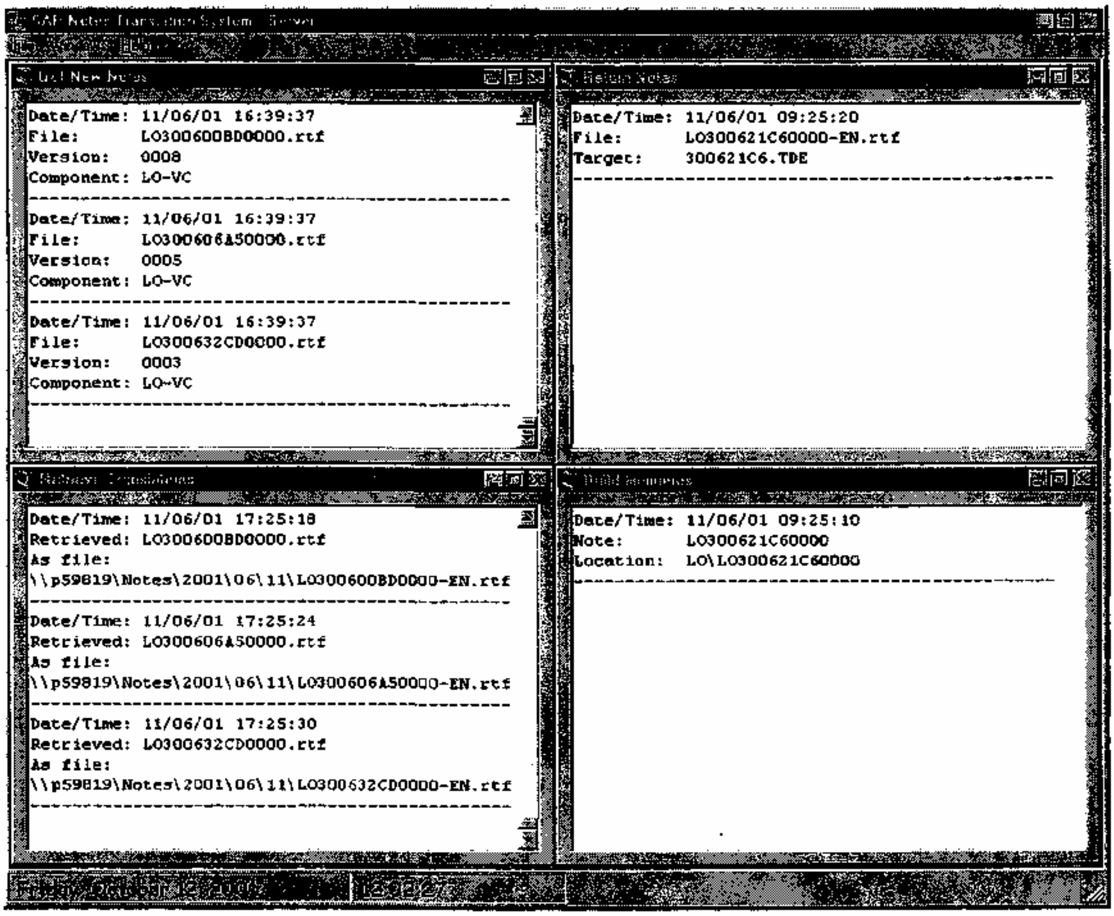
The Current Workflow Tool

The Notes are initially created, modified and stored in a central SAP System, called CSN. A special program has been developed for the MT translation, so that a user-defined selection of Notes can be downloaded from the CSN system to various directories on a local (Walldorf) file server, depending on the agency that will be translating the Notes. Two files are created - a parameter file containing various information about the Note (author, SAP-specific component, version number, ...) and the Note file (in RTF format) itself.

For each translation agency, a daemon program called the Notes Transfer Program (NTP) scans this download directory (plus several others) for files that require transfer to the agency system for translation. When a Note is downloaded, the NTP first makes backup copies of the files, in case there are any problems during later stages of the procedure.

The backups are saved to a dedicated file server, the Central File Store (CFS). Once the backup has been saved, the NTP checks to see if this is a brand new Note, or whether a previous version exists and has already been translated. If so, translation memory files should already exist on the CFS. The source and parameter files are copied to the agency site (normally to a temporary storage location at the agency site) along with the memory files, if these exist.

From here, another program, the NTS Server 'collects' the Note, runs various technical checks on it and inserts information about it into a central database (the DayDB). This central database is used to store various information about the Note, which can then be used either to assist in the processing of the Note or to enable various statistical analyses for the Notes translation at a later stage.



```
Date/Time: 11/06/01 16:39:37
File:      LO300600BD0000.rtf
Version:   0008
Component: LO-VC
------------------------------------------------
Date/Time: 11/06/01 16:39:37
File:      LO300606A50000.rtf
Version:   0005
Component: LO-VC
------------------------------------------------
Date/Time: 11/06/01 16:39:37
File:      LO300632CD0000.rtf
Version:   0003
Component: LO-VC
------------------------------------------------
```

```
Date/Time: 11/06/01 09:25:20
File:      LO300621C60000-EN.rtf
Target:    300621C6.TDE
------------------------------------------------
```

```
Date/Time: 11/06/01 17:25:18
Retrieved: LO300600BD0000.rtf
As file:
\\p59819\Notes\2001\06\11\LO300600BD0000-EN.rtf
------------------------------------------------
Date/Time: 11/06/01 17:25:24
Retrieved: LO300606A50000.rtf
As file:
\\p59819\Notes\2001\06\11\LO300606A50000-EN.rtf
------------------------------------------------
Date/Time: 11/06/01 17:25:30
Retrieved: LO300632CD0000.rtf
As file:
\\p59819\Notes\2001\06\11\LO300632CD0000-EN.rtf
------------------------------------------------
```

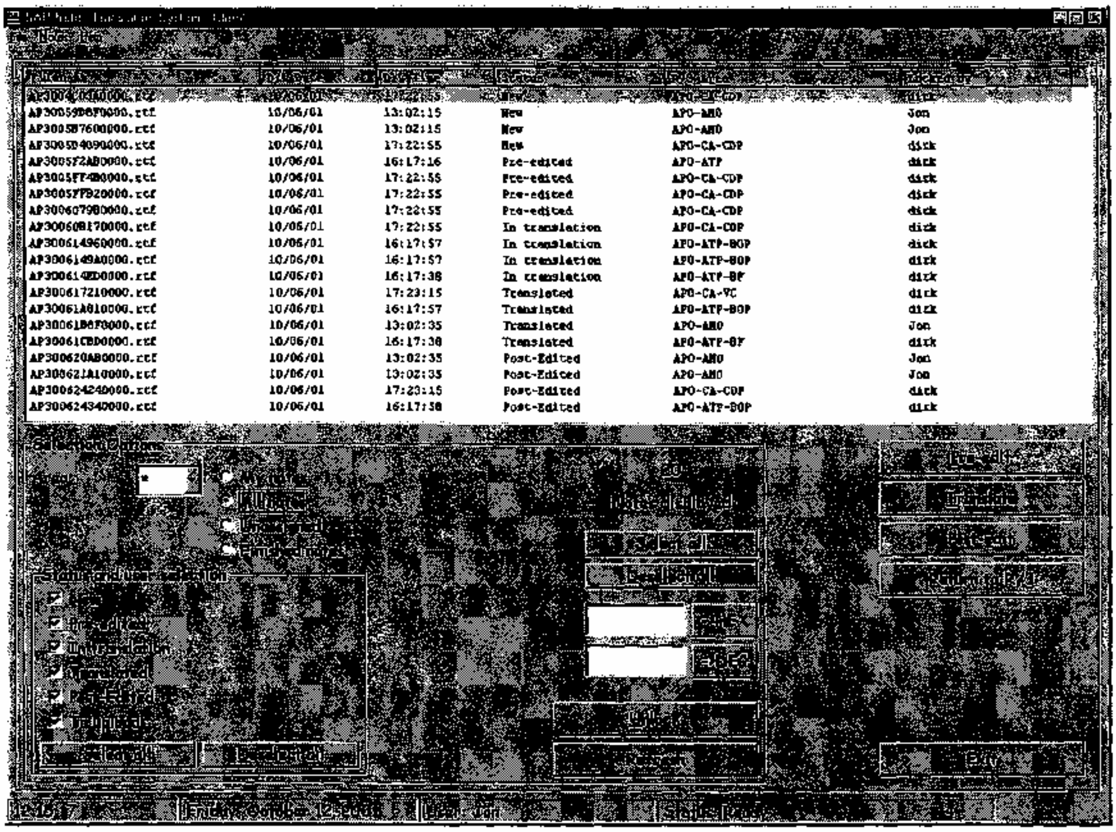
```
Date/Time: 11/06/01 09:25:10
Note:      LO300621C60000
Location:  LO\LO300621C60000
------------------------------------------------
```

Figure 2: The NTS Server

Figure 2 shows the NTS Server - the individual windows are processed in anti-clockwise order, starting in the top left. The window contents show various details (Note name, time and date stamps, file paths) about each Note that is processed by that particular step. There are four steps in total that are covered by the NTS Server and which run at various points during the translation procedure:

- *Step 1*
  Check downloaded Notes and insert them into the DayDB ready for processing.
- *Step 2*
  Retrieve completed translations from the MT system.
- *Step 3*
  Delete old memory files and create new ones
- *Step 4*
  Return Note to central directory for retrieval to Walldorf

Once inserted into the DayDB, with a status *New*, the Note then appears in the NTS Client - the software that the translators use for their day-to-day work. This is shown in Figure 3.



AP3004C0DA0000.rtf	10/06/01	17:22:55	New	APO-CA-CDP	dirk
AP30059D8F0000.rtf	10/06/01	13:02:15	New	APO-AHO	Jon
AP3005B7600000.rtf	10/06/01	13:02:15	New	APO-AHO	Jon
AP3005D4090000.rtf	10/06/01	17:22:55	New	APO-CA-CDP	dirk
AP3005F2AB0000.rtf	10/06/01	16:17:16	Pre-edited	APO-ATP	dirk
AP3005FF4B0000.rtf	10/06/01	17:22:55	Pre-edited	APO-CA-CDP	dirk
AP3005FFB20000.rtf	10/06/01	17:22:55	Pre-edited	APO-CA-CDP	dirk
AP30060798000O.rtf	10/06/01	17:22:55	Pre-edited	APO-CA-CDP	dirk
AP300608170000.rtf	10/06/01	17:22:55	In translation	APO-CA-CDP	dirk
AP300614960000.rtf	10/06/01	16:17:57	In translation	APO-ATP-BOP	dirk
AP30061494A0000.rtf	10/06/01	16:17:57	In translation	APO-ATP-BOP	dirk
AP300614ED0000.rtf	10/06/01	16:17:38	In translation	APO-ATP-BF	dirk
AP300617210000.rtf	10/06/01	17:23:15	Translated	APO-CA-VC	dirk
AP30061A810000.rtf	10/06/01	16:17:57	Translated	APO-ATP-BOP	dirk
AP30061B8F0000.rtf	10/06/01	13:02:35	Translated	APO-AHO	Jon
AP30061CBD0000.rtf	10/06/01	16:17:38	Translated	APO-ATP-BF	dirk
AP300620AB0000.rtf	10/06/01	13:02:35	Post-Edited	APO-AHO	Jon
AP300621A10000.rtf	10/06/01	13:02:35	Post-Edited	APO-AHO	Jon
AP300624240000.rtf	10/06/01	17:23:15	Post-Edited	APO-CA-CDP	dirk
AP300624340000.rtf	10/06/01	16:17:38	Post-Edited	APO-ATP-BOP	dirk

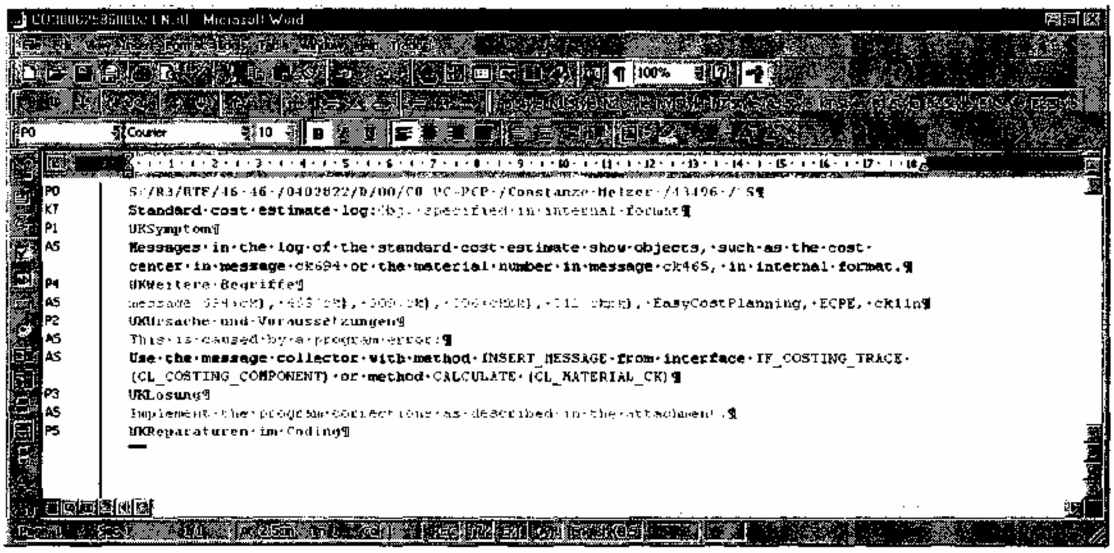
Figure 3: The NTS Client

The NTS Client enables the translator to select one or more Notes to work on and to then perform one of 4 steps on this selection of Notes, assuming they have a status that allows this. The various permissible statuses can be seen in Figure 3.

- **Pre-edit**
  Opens each file individually in Microsoft Word and runs several automatic checks on the files to ensure that they are correct. The template lines (lines that define and separate the individual sections of the Notes, see Figure 4) are marked in a red, bold typeface to indicate to the translators that they should not be touched and are marked as NoProofing in Word to indicate to the MT system that they should not be translated. The translator can then make any formatting changes to the Notes that he or she deems necessary to optimise the MT process, save the file and proceed to the next one. The DayDB status is then set to *Pre-edited.*
- **Translate**
  This function submits the selected file(s) to Comprendium for processing with MT/TM. If a TM already exists for this Note, it is also passed to Comprendium. The MT software first analyses the Note against the TM, taking only 100% matches into consideration, extracts the delta and runs it through the MT process. The resulting file can be seen in Figure 4. The NTS Server regularly checks the MT system to see if there are any files for which the MT/TM part of the translation procedure is complete and retrieves these, setting the DayDB status for the Note to *Translated.* Segments that have been extracted from the TM are coloured olive green, unknown words and constants

are also marked (red, purple) and the remainder (MT translations) is normal text (black).

- **Post-edit**
  Once the MT process has run, the translator has to post-edit the raw MT output and correct any errors. Again, MS Word is used and the selected Notes are combined into a single Master Document for simpler processing. The source and target languages are displayed side-by-side, although they cannot currently be scrolled simultaneously.
- **Return to R/3**
  The final step in the process is to return the Notes to Walldorf. This step actually triggers two separate processes. Firstly, new TM files are created from the source and target files, replacing the old ones. Secondly, the target files and the resulting memory files are subjected to various technical checks and copied to the local (agency) file store ready for transfer back to Walldorf. Note that the term R/3 is no longer used by SAP, but is used here, as it is still part of the user interface.



```
P0   S:/R3/RTF/46·46·/0402822/D/00/C0 PC-PCP·/Constanze·Melzer·/43496·/·S¶
KT   Standard·cost·estimate·log:Obj.·specified·in·internal·format¶
P1   UKSymptom¶
AS   Messages·in·the·log·of·the·standard·cost·estimate·show·objects,·such·as·the·cost·
     center·in·message·ck694·or·the·material·number·in·message·ck465,·in·internal·format.¶
P4   UKWeitere·Begriffe¶
AS   message·694·(CK),·465·(CK),·500·(CK),·106·(CKMR),·111·(CKCS),·EasyCostPlanning,·ECPE,·ck11n¶
P2   UKUrsache·und·Voraussetzungen¶
AS   This·is·caused·by·a·program·error:¶
AS   Use·the·message·collector·with·method·INSERT_MESSAGE·from·interface·IF_COSTING_TRACE·
     (CL_COSTING_COMPONENT)·or·method·CALCULATE·(CL_MATERIAL_CK)¶
P3   UKLosung¶
AS   Implement·the·program·corrections·as·described·in·the·attachment.¶
P5   UKReparaturen·im·Coding¶
```

Figure 4: A Typical Note After Post-Editing

In the last stage of the translation process, the NTP scans the agency file store for any files that are being returned by the agency and returns these to a special upload directory within Walldorf, where they are collected by the CSN system and made available to SAP customers worldwide.

Current Restrictions

The system as currently implemented exhibits a number of restrictions. Some of these, however, have only come to light as the system has grown and developed. Fortunately, none of them have yet caused major disruptions to the translation process...

- **Scalability of NTS Databases**
  The NTS is based on Microsoft Access databases. This is an excellent tool for rapid implementation and for small groups of users, but has limitations - once

20 or more users are connected to a database, the performance begins to suffer and by the time roughly 30-35 users are connected, the database becomes virtually unusable.

- **Data Storage Bottlenecks**
  All of the translated material, as well as the memories and various administrative files must be stored for future retrieval and/or analysis. This data store currently contains approximately 3.5 million files, and is growing at the rate of 7,500-10,000 files per day. This already causes major problems with data backup and retrieval solutions and may also soon pose problems for the Windows operating systems themselves.

- **Downtime**
  If any part of the chain of servers and programs between SAP and the translation agencies is broken (any of the SAP servers involved, the Notes Transfer Program, the connection itself or any part of the agencies hardware or software), the entire NTS is rendered more or less useless. If an extended period of downtime occurs, the translators quickly run out of work to do and cannot access any more until the system has been fully restored.

- **Restrictive System Design**
  The current NTS does not easily enable any other MT or TM systems to be used, and is limited to translation DE-EN. All of this is currently hard coded.

- **Memory Restrictions**
  Currently, the TM system only allows for 100% matches and does not enable any interaction between the user and the TM. There is one TM for each Note, but no cross-reference between related Notes. Any similar or even identical segments in other Notes cannot be taken into consideration during the translation procedure.

Future Plans

In order to be able to continue using the NTS and to keep up with the constant increases in the numbers of Notes to be translated, a number of modifications are going to be required in the near future. Currently, the following plans are either being discussed or are already being implemented to ensure system stability and usability in the future.

Background Memories
Improvements in API availability and performance have drastically changed TM systems over the last few years. We are planning to implement improved memory handling (probably using TRADOS 5), allowing for interactions between the user and the TM system as well as the use of large background memories with fuzzy logic. This will increase the match rate during the translation procedure, thus further reducing the costs for SAP.

Modularisation of the NTS
This would allow other MT and TM systems to be more simply integrated with the NTS workflow tools and other language pairs/directions to be implemented.

Eventually, other text types could also be translated using the same set of tools. EN-JA is currently in development at SAP.

Better DB scalability
The DB at the heart of the NTS will have to be changed. Likely candidates are currently SQL Server or SAPDB (previously ADABAS). This will provide faster, more stable and more scalable access to the central information within the NTS.

Better Integration with the SAP System
The SAP System has many features, particularly the integrated translation tools, which are specifically designed to handle mass data and could be used to improve and streamline the entire workflow process. The simple file download may be replaced with more complex modules, possibly replacing some of the functions of the NTS tools.

DB-based file storage
This concept is very similar to that of a "traditional" document management system, where files are checked in and out of a database when used. This immediately removes the large number of individual files that the operating system and backup solution have to cope with.

Downtime reduction
Possible solutions here include a simple system redundancy (duplicate systems, servers, connections, etc.), but will probably also include building a more intelligent solution which is capable of automatically preparing (and, where necessary, refreshing) a buffer of work at the agency site, thus reducing the effect of any downtime on the part of either SAP or the telecommunications companies.

Glossary of Abbreviations

API	Application Program Interface
CFS	Central File Store
CSN	SAP System for Internal and Customer Support
DB	Database
MLT	Multilingual Technology (Group at SAP AG)
MT	Machine Translation
NTP	Notes Transfer Program
NTS	Notes Translation Solution
TM	Translation Memory

References

ASLIB Proceedings, 1998
Paper presented by C. Pyne & D. Grasmick

MT Summit Proceedings, 2001
Paper presented by J. Brundage