ELSEVIER

Computer Networks and ISDN Systems 30 (1998) 1237–1248

COMPUTER NETWORKS and ISDN SYSTEMS

Twenty-One: cross-language disclosure and retrieval of multimedia documents on sustainable development

W.G. ter Stal [a], J.-H. Beijert [a], G. de Bruin [a], J. van Gent [c,*], F.M.G. de Jong [b], W. Kraaij [c], K. Netter [e], G. Smart [d]

[a] *Getronics Software, P.O. Box 22678, 1100 DD Amsterdam, Netherlands*
[b] *University of Twente, Faculty of Computer Science, P.O. Box 275, 7500 AE Enschede, Netherlands*
[c] *TNO–TPD, Multimedia Group, P.O. Box 155, 2600 AD Delft, Netherlands*
[d] *Highland Software Systems, P.O. Box 2035, 2002 CA Haarlem, Netherlands*
[e] *DFKI, P.O. Box D-66123, Saarbrücken, Germany*

Abstract

The Twenty-One project brings together environmental organisations, technology providers and research institutes from several European countries. The main objective of the project is to make documents on environmental issues—in particular, on the subject of sustainable development—available on CD-ROM and on the Internet. At present, these documents exist on different media (paper, electronic documents, audio-visual material), in different formats (HTML, word processor) and in different languages. This diversity impedes the distribution of documents through normal channels, and makes it hard to search for and retrieve targeted material on specified subjects. The project is developing search engines that can locate required information, and uses automatic translation tools to make foreign-language texts available. Authors and editors gain economic advantages by the increased distribution of their documents, and users find documents more readily available and easily accessible. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Multimedia handling; Cross-language information retrieval; Information transaction model; Environmental information exchange

1. Introduction

1.1. Objectives of project Twenty-One

The main objective of Twenty-One is to develop domain-independent technology to improve the quality of electronic and non-electronic multimedia information, and make it more readily and cheaply accessible to a large group of people. At the outset, Twenty-One will prove this technology in the field of ecology and sustainable development. However, the generic characteristics of the distinct software modules allow for easy application outside the domain of environmental information exchange.

The technology developed by Twenty-One facilitates access to information by readers who are not native speakers of the language in which the information is provided. The project result will be a software demonstrator that enables users to type in queries in one of four selected European languages specifically Dutch, English, French and German and to retrieve (multimedia) documents. The core of the

[*] Corresponding author: E-mail: gent@tpd.tno.nl

0169-7552/98/$19.00 © 1998 Elsevier Science B.V. All rights reserved.
PII: S0169-7552(98)00004-X

system consists of an index in the four languages that has been built and translated automatically by the software. The envisaged demonstrator will further allow users to communicate interactively with the providers of the documents. The demonstrator will use both CD-ROM and the Internet and its de facto applications (such as WWW, E-mail, etc.) as media. The periodically distributed CD-ROM is used for rapid access to static document bases, whereas Internet will be used for dynamic data and document bases. In addition, the Internet will be used for interactive communication with all parties involved in the dissemination and the transaction model.

The Twenty-One information transaction model, also called the Galilei model, forms an important prerequisite to employing the technology developed within the project. The information transaction model allows different environmental organisations to exchange that is, publish and retrieve different information objects. As such, the information transaction model can be view as a sociotechnical system. This term originates from organisation theory [14] and refers to systems where emphasis is laid on the interdependencies between technical equipment and the (groups of) people using this equipment.

Although the information transaction model constitutes an important framework for the success of project Twenty-One, the main emphasis in this paper lies on the technology designed for efficient and effective document disclosure and retrieval.

1.2. Project organisation, context and time schedule

Twenty-One is a project funded within the EU Telematics Applications Programme, sector Information Engineering. Project partners are academic, the University of Twente and the University of Tubingen and commercial software companies, namely Highland Software Systems (UK) and Getronics Software (NL). In addition, Twenty-One includes contract research organizations, namely DFKI (D), Rank Xerox Research Center (F), and TNO–TPD (NL) together with a number of non-profit environmental organizations, Environ Trust (UK), Friends of the Earth (B), Klima-Bündnis (D), MOOI Foundation (NL), VODO (B). The name ‘‘Twenty-One’’ refers to Agenda 21, the document resulting from the UN conference on sustainable development in Rio de Janeiro in 1992. The duration of the project is 36 months and the starting date was March 1996. According to plan, the project launches its main software products in July 1998. As a consequence, this paper reflects work in progress.

1.3. Overview of remaining sections

In Section 2, we briefly describe the motivation and key features of the information transaction model that forms the backbone of the Twenty-One approach. Section 3 contains an overview of the applied technology. Related technologies, exploitation and further plans are discussed in Section 4. In Section 5, we summarise the main aspects of project Twenty-One.

2. The Galilei model

The environmental partners within Twenty-One have developed an information transaction model, the Galilei model. Both information providers and seekers profit from the model, the former by increasing the number of potential customers, the latter because more information becomes available. The project supports the objectives of the users involved in the project by trying to stimulate interaction and raise awareness of local Agenda 21 initiatives in Europe.

2.1. Rationale for an information transaction model

Successful application of the software developed in the Twenty-One project within the field of sustainable development will greatly depend on the willingness of key players to participate in the information transaction. When producers, providers and publishers are not willing to provide their information, or information brokers are not willing to transfer the information, or when end-users are not expressing the need for information, nothing will flow through the information channels of Twenty-One. In order to get the different parties to participate in the information transaction, they will have to be motivated to do so. The Galilei model provides a solution for this problem, by making the advantages for all participators in the information transaction so clear that information will keep flowing by itself. The model will function as a perpetuum mobile, that is, information transaction will take place automatically, without incentives from outside. In this sense, the Galilei model resembles the economic market mechanism.

2.2. Key features of the Galilei model

The Galilei model has as its most important features:

- enlargement of the number of connections between the elements of the information domain to permit a richer flow of information;
- shortening of linkages to enable a fine-tuning between demand and supply and, consequently, a faster flow of information;
- enlargement of the number of players, and through this the enlargement of both demand and supply;
- the increase of the number of roles per player and, consequently, an increase of both demand and supply;
- regulations to ensure the autonomy of the players; by doing this the biggest obstacle to potential participants taking part in the information exchange has been removed;
- the offering of new perspectives to players; by this, they are stimulated once again to take part in the information exchange;
- the creation of new roles for the existing players and for new players, to support all processes;

3. Technical aims and achievements

This section contains an overview of the technical aims of project Twenty-One. As said in the introduction, this paper reflects mostly work in progress. However, some modules have been designed and implemented in previous projects, therefore it is already possible to point out some accomplishments.

The main technical product of project Twenty-One is the so-called Twenty-One Demonstrator. The basic functionality of the Twenty-One Demonstrator allows (a) end-users to get easy and cross-language access to a multilingual and multimedia information base, and (b) publishers to submit and disclose their information at very low costs.

The Twenty-One Demonstrator concerns software to be delivered by the end of project Twenty-One, including two crucial sets of software:

- Software to disclose multimedia information.
- Software to retrieve multimedia information (accessible with current state-of-the-art Internet browsing applications, such as Netscape) from remote servers or from a local CD-ROM. The core of the Twenty-One retrieval software [1] consists of a search kernel supporting several query modes and interface languages.

The Demonstrator's functions will be *disclosure*, *maintenance* and *retrieval* of multimedia information. By *disclosure* is meant the process of automated attachment of features to information objects so that they can be found. These features will in many cases consist of index terms. *Maintenance* is the set of procedures to keep the database consistent and up-to-date. This topic is not further addressed in this paper. *Retrieval* refers to the functionality of the system to find relevant information on the basis of a user's queries.

In Section 3.1, we discuss the design decisions forming the starting point of the software development within project Twenty-One. In addition, we provide a short functional description of the technical modules. Section 3.2 focuses on the disclosure part of the Twenty-One Demonstrator. The retrieval functionality of the Demonstrator is discussed in Section 3.3.

3.1. Design decisions and global system characteristics

3.1.1. Design guidelines and principles

At the outset of the project, we have taken a number of design decisions which should facilitate development of the project software and minimise budgetary risks. As such, the following design decisions constitute a framework within which the de-

[1] The Twenty-One retrieval functionality can be inspected through: http://twentyone.tpd.tno.nl/.

signers, engineers and representatives of the user group agreed to co-operate.

The design decisions currently employed in project Twenty-One amount to:

Reuse and extension of existing advanced technologies. This design decision applies to several levels of technology. At first, it holds for reusing distinct software modules originally designed and implemented by one of the consortium partners, and currently being integrated and /or extended within the overall architecture of the Twenty-One system. For instance, the Xerox tool kit Xelda, see Section 3.2.3, is a key module within Twenty-One, but has been developed outside the context of the project. Besides software that contributes to the core functionality of the system, we are employing existing, advanced software tools for planning, communication and Web interface design.

A generic and modular, component-based architecture. The fact that we adhere to a reuse and sharing policy as much as possible almost automatically yields a modular, component-based architecture. This is not only a matter of favourable design policy. Since the Twenty-One consortium consists of several commercial and research partners, one is confronted with proprietary software and license agreements. Therefore, the choice of a modular architecture, and the subsequent focus on designing and developing suitable interfaces, is inevitable.

Use of open platforms and de facto standards. The Twenty-One software is initially intended for a large number of users within environmental organisations across Europe. A user survey within this group revealed that the majority of these groups use the Windows 3.11 platform and a standard Internet browser, such as Netscape or Explorer. As a consequence, the first releases of the Twenty-One software are targeted to run on this platform. However, for other generic applications, most software is also available for Unix and Windows NT platforms. In any case, we presume a standard three-tier client/server architecture and the availability of standard auxiliary applications, such as word processing software, multimedia viewers and Email.

Incremental and co-operative design. Being a project with 12 institutions, geographically distributed over five countries, it is difficult to work towards a fixed design and through an implementation phase in a straight line. Therefore, we have planned and already launched several releases of (parts of) the Demonstrator, which can be commented on through the WWW. By means of checklists, the future end-users within the consortium can comment on new releases.

Evaluation planning and quality assurance. During the project, evaluation is being given careful attention. We have planned several internal activities for monitoring and evaluating the quality of the software modules. In addition, the project takes part in external benchmarking activities, as is shown by our participation in the TREC 6 Special Track on Cross-Language Information Retrieval [5].

3.1.2. Global system characteristics

The aforementioned design assumptions constitute the basics for the development of the Twenty-One Demonstrator, embracing disclosure as well as retrieval functionality. The global functionality of the modules can be summarised as follows:

Multimedia handling. The Twenty-One system aims at the disclosure of documents of different media types and/or data formats, e.g. paper documents, WEB documents, word processor documents, text annotated images, audio or video material.

Document conversion. The system incorporates a component for the conversion of the various document formats into standard representation (SGML/HTML), including a tool for the conversion of paper documents into electronic format, on the basis of layout analysis and OCR.

Advanced disclosure techniques. The Twenty-One Multimedia document base will be disclosed using several advanced techniques —like rule-based Natural Language Processing (henceforth: NLP) for phrase indexing, relevance ranking and automatic hyperlink generation.

Domain-tuning: sustainable development. The aim of the project is to build a system that supports and improves dissemination of information about ‘‘local Agenda 21’’ initiatives. This requires a special effort in the acquisition of linguistic resources that are tuned to the language and vocabulary in this domain. The technology to be developed is still supposed to be generic.

Targeted at various publishing media. The disclosure system produces an index on a multilingual multimedia document base. Indexes built by the system can be made available via CD-ROM, or they can be made accessible via a Web-server.

Cross-language retrieval / multilinguality. Twenty-One offers the possibility to retrieve documents in another language than the query language. The languages presently covered in the project are Dutch, English, French and German; extension to other languages is being considered.

Automatic hyperlinking. The automatic hyperlinking function attaches typed hyperlinks between terms, phrases, or images, etc. These links can be either static (generated off-line) or dynamic, in which case links are attached to pages at run-time by a CGI-program.

In the next section, we zoom in on the disclosure modules within Twenty-One.

3.2. Disclosure of multimedia documents

An important prerequisite for information retrieval, or information searching, is that information objects are identified and disclosed. The term *disclosure* refers to the assignment of features to information objects such that they can be located. As such, disclosure not only refers to the process of indexing documents, it also involves, as will explained in the remaining sub-sections, scanning, optical character recognition (OCR) and document enhancement.

It is well known that manual disclosure of information is costly. Specifically, the costs of manually attaching features to information objects tend to be enormous. In addition, disclosure of information by humans often yields inconsistencies and other errors, which prevent an optimal retrieval process.

In order to overcome these deficiencies within Twenty-One, we designed, combined and re-used several disclosure modules which render a text-based index in an automated fashion. To date the disclosure modules of Twenty-One are capable of handling the following multimedia information objects:

- Facsimile (scanned) images of the sheets of paper on which the text was originally published;
- Electronic text documents in several (file) formats, including HTML pages located on remote websites;
- Line drawings and graphs;
- Halftone images (photographs);
- Soundtracks;
- Video and film fragments.

Fig. 1 shows how paper documents and other media objects are processed and disclosed for inclusion in the Twenty-One database. There are two main information channels through which data flows into the Twenty-One database.

(1) Paper documents are scanned and disclosed using Layout Semantics Discovery (LSD) and optical character recognition (OCR). The results of OCR are then indexed into the Twenty-One database as text records. This process is described in Section 3.2.1.

(2) Multimedia objects are captured and digitized using the Multimedia Editor, and the resulting files are linked, using the editor again, to text records in the database. The different kinds of multimedia objects are captured by the Multimedia Editor in different ways: paper-based objects, such as linework and photographs, are scanned and captured by document enhancement; audio and video objects are clipped and reformatted; word processor and other files are converted. See also Section 3.2.2.

The remaining sections are organised as follows. First we will describe in Section 3.2.1 how paper documents are being processed. In Section 3.2.2 we explain how multimedia objects (images, video fragments, sounds) can be incorporated or linked to the Twenty-One document base. The construction of the phrase-based index and its translation are discussed in Section 3.2.4.

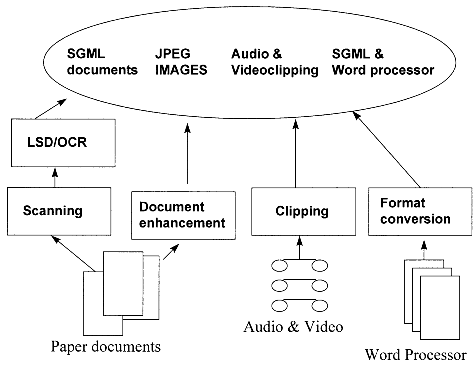


Fig. 1. Processing and disclosure of different information objects for inclusion in the Twenty-One database.

3.2.1. Scanning, OCR and Layout Semantics Discovery

Scanning is the process of converting images on paper into digital files. The Twenty-One project will use a high-volume monochrome (black-and-white) scanner, with sheet-feeder, to support fast batch scanning of document pages. The input consists of paper (A4) pages, from which the scanner generates TIFF images. Subsequently, TIFF images are fed into a process of Layout Semantics Discovery and OCR. The two processes of Layout Semantics Discovery and OCR have been integrated into one functional module of the Twenty-One demonstrator.

Layout Semantics Discovery (LSD) is a method for the automatic extraction and tagging of logical components in a document. Examples of logical components are headings and paragraphs. The purpose of LSD is to impose on each document an SGML mark-up according to the appropriate Document Type Definition (DTD) that describes the logical structure of documents of this kind. The LSD process consists of a sequence of image-processing steps, that extracts connected components and their features, followed by a sequence of labeling steps, in which different components are grouped into high-level objects and tagged as SGML elements.

Once document components have been tagged, they can be converted to ASCII using OCR/ICR software. Optical character recognition (OCR) and intelligent character recognition (ICR) refer to the process of converting bitmap representations of characters into ASCII-coded text that can be read and edited by a word processor. There are many commercially-available OCR packages, and Twenty-One has decided to use Xerox's TextBridge™, one of the three [2] leading commercial OCR systems.

3.2.2. The Multimedia editor

The main functionality of the Multimedia editor is to (re)scan, digitize or to convert different types of multimedia information, and to allow a user to link this information manually to relevant text records in the Twenty-One database. Currently, the Multimedia editor supports, besides the linking capability, three main processes, (1) *document enhancement*, (2) *audio and video clipping* and (3) *format conversion*.

3.2.2.1. Document enhancement. Some multimedia objects are physically associated in the input information source with textual information, in that they are originally published on the same paper page as the text. Examples of such objects are linework (graphs and drawings) and halftones (photographic images). Although linework is normally originated as vector drawings using programs such as CAD packages or relational database systems, in the Twenty-One system all document illustrations, linework and halftones will be captured and stored as bitmaps. This will include other objects which contain text, but where the semantic content of the object resides in the way the text is arranged: for example, complex tables will be stored as bitmaps.

Document texts are enhanced, in Twenty-One, by extraction and storage of these multimedia objects from the document, including rescanning when required. It is technically possible to scan a document page in such a way that the text can be extracted from the scan for OCR and indexing, while the artwork can be processed from the same scan into separate bitmap files for storage. But this is not the best way to perform data capture. Text pages for OCR can be batch scanned at high speed for OCR, using a binary (black-and-white) scanner, whereas photographs require hand processing using a slow, high-quality color scanner, usually with adjustment of the settings for each picture to compensate for variations in brightness, contrast, color density, and so on.

Document enhancement therefore involves a second scanning process, using a high-quality hand-fed color scanner, to capture illustrations from the paper page, or, when available, to scan the original artwork or photograph used for the original publication. Illustrations that are captured on the first scan, and recognized by LSD as non-text objects, will be stored as JPEG files following LSD/OCR. These files can be replaced by high-quality bitmaps obtained during secondary scanning and document enhancement.

[2] The other two are WordScan from Calera, and Omnipage from Caere Corporation. Each year the Information Science Research Institute (ISRI) at the University of Nevada in Las Vegas tests the accuracy of various OCR packages, and these three OCR packages consistently perform equally well.

3.2.2.2. Audio and video clipping. Besides artwork and photographs, other multimedia objects (such as sound or video sequences) can be available for disclosure. Although these can be associated semantically with text records to be stored in the Twenty-One database, they are published on different media, and their only association with the original printed page will be a text cross-reference.

Capture of these multimedia objects involves digitizing the information contained on an analog storage medium, such as an audio-cassette tape or a videotape, and then reformatting and compressing the resultant file(s) to the parameters required for the Twenty-One database. Sound and video files, even after compression, are very large. If they are to be stored in a database and transmitted across a network, their size must be constrained by setting limits on sound and picture quality, picture size and frame rate.

The process of clipping comprises the digitizing and reformatting of sound and video-tapes. Output from the process will be a set of files in standard formats (WAVE files for sound, MPEG-1 files for video) for attachment, by the Linker, to text records in the Twenty-One database.

3.2.2.3. Format conversion. Text information will be available for inclusion in the Twenty-One database from sources other than scanned document pages. Files from word-processor systems, desk-top publishing systems, spreadsheets, and so on, will contain text that can be indexed and stored in Twenty-One. Format conversion is the process of converting files, such as these, into SGML-marked text files.

Ideally, the word processor, DTP package or spreadsheet will itself be used to generate an SGML or HTML file, or at least an ASCII text file. If this is not possible, format conversion tools will be used to extract recognizable text fields from the file, remove control characters, and insert SGML tags. The different formats to be supported for conversion will depend on the document corpus collected for the Twenty-One demonstrator, and on its content, but format conversion will at least support MS Word 6.0 and MS Excel 5.0 formats.

As a result of document enhancement, audio and video clipping, and format conversion, the Multimedia Editor buffer will contain files that require to be attached to text records in the Twenty-One database. This attachment is performed by (environmental) information specialists (information brokers) using the Linker component in the Editor.

3.2.3. Phrase-based indexing using NLP

The link between the disclosure and retrieval modules is formed by the automatically acquired text-based index. The Twenty-One index is, unlike most ordinary retrieval systems, not limited to an inverted file index based on single words or lemmas. It combines phrase based indexing with a vector space model (VSM). Experiments with phrase-based indexing within Twenty-One have shown a considerable improvement in retrieval performance (see Refs. [6–10]). Using a phrase-based index, users are allowed to query the system by using complete phrases, extending the level of simple keywords, such as: *effects of acid rain on forests in the Netherlands.*

Phrases—potential index terms—are identified during the disclosure process by the NLP modules working on the HTML/SGML documents originating from the Twenty-One document base. See also Fig. 2.

NPs will be extracted automatically from the texts using language detection, grammatical analysis, heuristics and statistics. Grammatical analysis will take place in two stages:

(1) tagging: the process of attaching features to words in texts, using lexical and contextual information;

(2) parsing: combining words into structural units or phrases.

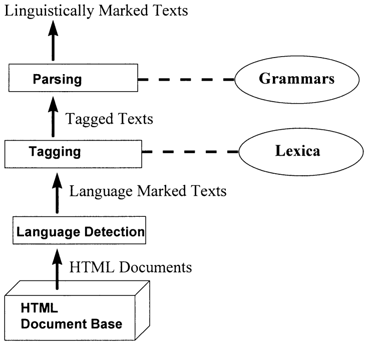


Fig. 2. Noun phrase extraction from the Twenty-One document base.

3.2.3.1. Tagging. The general purpose of a part-of-speech disambiguator (*tagger*) is to associate each word in a text with its morpho-syntactic category (represented by a tag), e.g. to determine if, in a given sentence, a word is a singular noun, a finite verb, a preposition, etc. In Twenty-One, tagging will be based on the Xerox Linguistic Development Architecture (XeLDA). The Xerox taggers rely on the same architecture for the different project languages, which include Dutch, English, French and German. Unknown words are handled by a guesser, which specifies potential part-of-speech tags based on specific affixes. Taggers are generally not fully accurate. In most cases, the error rate ranges between 3 to 5% of the words which are improperly tagged. After word tagging, the parser will identify the phrases in the texts.

3.2.3.2. Parsing. The parsing component assigns a structural analysis to the text for the purpose of identifying noun phrases (NPs). These NPs form the basis for indexing. Within Twenty-One, there are two types of parsers available, a deterministic LR parser and a non-deterministic Tomita parser (see Ref. [12]). Both parsers will operate on the same context-free, phrase structure grammars for the languages involved. Both parsers can be generated by a parser-generator contained in the TNO NLP toolkit. The main features of the TNO parser-development kit include:

- declarative grammar specification
- simple grammar formalism, easy to handle for a grammar writer
- parser speed
- parser robustness

The parser development kit consists of a parser generator, which takes a phrase structure grammar as input and produces a LR grammar specification. This LR grammar can be compiled into a deterministic parser by the use of YACC [3]. In the same manner, a chart parser [12,13] can be produced by using NLY-ACC [4]. The deterministic version cannot handle ambiguous tokens. The non-deterministic version can handle lexical and structural ambiguity at the cost of some speed.

The parsing component will preserve the HTML document mark-up for later stages of processing: this is essential for the support of hyperlinking across documents (see also Section 3.3).

3.2.4. Partial translation of Twenty-One document base

After identification, NPs can be submitted to the term translation (henceforth, TT) module. By terms we refer to the main indexing units within Twenty-One: NPs. In most cases, a term is complex, i.e. consists of more than one word. The challenge is to develop robust term translation techniques which can preserve the morpho-syntactic information of the NP structure. This structure is available because every document is processed by the monolingual NLP modules, as described in Section 3.2.3. Identification and translation of multi-word expressions is a well-known problem [11], but, by combining corpus based approaches and bilingual dictionaries, this problem can be tackled up to a level that is adequate for the purposes of cross-language information retrieval (CLIR). In a nutshell, TT fulfils three roles:

(a) It is the basis for the generation of a set of monolingual indexes (one for each project language). The monolingual NLP modules identify the NPs in a document as the indexing units. By off-line TT, these source language index terms receive three target language equivalents. These index terms are stored in the four monolingual indexes. During retrieval, queries are matched with these monolingual indexes.

(b) If monolingual query handling does not lead to any hits, TT can also be applied on-line to the translation of query terms. Query translation can partly alleviate the effects of poor quality MT in the following ways:

1. A document with a relevant term, which contains an OCR error, can be found via fuzzy matching with the translated query.
2. The user can perform relevance feedback in the target language, once a relevant document is found in the particular foreign language. This technique is also useful to overcome the effects of translation ambiguity.

[3] Cf. the manual pages of any UNIX system.

3. A word-based translation approach followed by a ranked Boolean query (cf. Ref. [2]) can act as a disambiguating filter.

(c) TT is the basis for the establishment of hyperlinks between terms and their translations. The result is a (part of a) document, aligned with its three translations. The alignment between terms will be implemented by hyperlinks.

In addition to partial translation by means of TT, we are currently experimenting with off-line Document Translation (DT) for the purpose of enabling users to judge the relevance of retrieved material. Experiments are planned with full-text translation by the on-line software made available by LOGOS. Full-document translation could also be used as a basis for monolingual indexing in the three target language versions of a document. A disadvantage of DT systems is that they are file-oriented and thus require post-translation alignment (reverse engineering).

Both NLP and TT require lexical resources. Machine-readable dictionaries, as owned by commercial lexicon publishers, could be useful for the generic lexical knowledge required by the monolingual NLP-components and the translation modules. Such lexical databases usually not only contain information on single words, but also contain idioms and collocations plus their translation, which can be extremely valuable. However, the application of machine-readable dictionaries for our purposes is complicated by the fact that coverage of all the four project languages is rare. Therefore tools that can automate the acquisition of lexical resources are not only important for the domain specific vocabulary, but could also be of value for the generic part. The dictionaries envisaged for Twenty-One will merge these various lexical resources. For an overview of selected techniques and experiments, see Refs. [2,3].

3.3. The Twenty-One search kernel

In order to assure maximal flexibility, the Twenty-One retrieval module supports several query- and search strategies. In this section, we briefly discuss the main characteristics of the search and retrieval components. In Section 3.3.1, emphasis is put on the global functionality. Section 3.3.2 discusses some issues in designing the Twenty-One retrieval interface.

3.3.1. Retrieval functionality

3.3.1.1. Two-step retrieval strategy. Typically, a retrieval session takes place in a two-step fashion. First a user enters either a free text (words or phrases), Boolean or bibliographical query. In case of a free text query, documents are ranked on the basis of the similarity and number of matching phrases. The fuzzy matching component will enable retrieval of morpho-syntactic variants and phrases mutilated by OCR errors. The free text query mode enables a user to search with high precision. In case the user identifies a page as being particularly relevant, he, or she, may decide to feed this entire page back into the retrieval system using the VSM index. This latter search mechanism is known as *relevance feedback* or *search similar*. The rationale for this two-step approach is that, by means of a first search, the user is led to potentially interesting documents, that in a next step may be used to improve retrieval results.

3.3.1.2. Typed hyperlinking. Another feature of the Twenty-One retrieval interface is that it allows the inspection of information objects by clicking on so-called *typed hyperlinks* (see Ref. [1]), which are visible at the retrieved document page(s). The Twenty-One hyperlink searching follows the de facto standards available on the Internet, with the exception that, in Twenty-One, hyperlinks are typed. If a user selects an object (usually a term) that has a link to another object or a set of objects, the user will be informed about the ‘‘type’’ of the links. Examples of object link types are:

Translation: selection of an object with a link of type ‘‘translation’’ yields a list of languages in which the system contains a translation of the term.

Definition: the link type ‘‘definition’’ points to a list of text objects where the user can find a definition of the term.

Similar occurrence: selection of an object with a link type ‘‘similar occurrence’’ provides a list of documents or objects in which a morpho-syntactic variation of the term can be found.

Picture: this link points to objects involving a picture with a caption that contains (a morpho-syntactic variation of) the term.

Compared to standard hyperlinking, the employment of typed hyperlinking results in a faster and a more informed inspection of additional or supporting information objects.

3.3.1.3. Cross-language information retrieval. Having selected one or more parts of the Twenty-One document base in a particular language, the user is able to retrieve relevant documents besides the one available in his or her native language. Although documents are, when only using TT, partially translated, the idea is that generally they provide enough clues for a user to judge whether the documents are relevant for a specific information need. If a full translation is desired, the document may be sent to an on-line MT service, for instance LOGOS, or to a standard translation service.

Multilinguality is also supported by means of a language switch that allows users to select the preferred interface language, that is the language in which all headers, footers, buttons, check boxes and help texts is displayed.

3.3.2. Designing the retrieval interface

The design of the Twenty-One user interface follows the functionality of the underlying system, the technical possibilities and limitations, and the wishes of the users of the interface. Some of the issues relevant for Twenty-One are discussed below.

3.3.2.1. Functionality issues. The Twenty-One Demonstrator is a search engine and the design of the interface follows the standard build-up of such engines on the Internet: a screen for query entry, the presentation of a list with results ordered by relevance, and the presentation of a selected item from that list. A distinction between the normal search engine and the Twenty-One retrieval system is the two-step retrieval strategy (see Section 3.3.1). This extra feature must be stressed in the design of the interface and be accessible by the press of one button.

The Twenty-One Demonstrator is multilingual. The interface itself is presented in one of four languages, the query can be in any of the supported languages, and all documents can be viewed in any of the languages for which a translation is available. The interface must offer a default language for all presentations and an easy way to switch between languages. When presenting the documents, it must be clear what the original language of the document is.

The presentation of the documents is enriched with typed hyperlinks (see Section 3.3.1). In theory, one word can link to multiple types of references. The interface must indicate what kind of hyperlinks are available on a page, and it must be clear what the result of clicking a link is. This is done by adding buttons that are highlighted when a type of hyperlink is present on a page. By pressing the highlighted button, the hyperlinks are made visible on the screen, and can be clicked like standard hyperlinks.

When using the Twenty-One retrieval system, users must be made aware of the fact that they can contribute to the database by adding new material, or by commenting on the available documents (the Galilei model, see Section 2). The interface must lead the user through the steps needed to submit new material, comments and ratings of documents.

3.3.2.2. Technical issues. The design of user interfaces for the WWW is limited by the capabilities of browsers, which, in their turn, are limited by the capabilities of the system on which they run. In the case of Twenty-One, some of the users have a slow connection to the Internet and low performance systems. This means the interface must be built with standard HTML, without enhancements like Java applets. Due to limited bandwidth, the use of graphics must be kept to a minimum.

3.3.2.3. User involvement. One benefit of building an interface for the World Wide Web is the possibility of involving the users in the design via the Internet. During the development of the interface, a series of prototypes are made available to the users, giving them the opportunity to look over the shoulder of the interface designer. The users can give comments, using a questionnaire that is included in the prototype interface. These comments are incorporated in the next version of the prototype, which can be evaluated by the users again. In this way, the final interface design is the result of an incremental and cooperative process.

4. Related technologies, exploitation and future plans

The Twenty-One technology, as discussed in Section 3, has been designed for efficient and effective document disclosure and retrieval. Further enhancement of this ‘‘product’’ could be achieved by relying on emerging technologies, such as image recognition, multidimensional graphic display of information spaces based on 3-D technology, and so-called intelligent agent technology. The Twenty-One technology is widely applicable. The software packages Twenty-One Publisher and Twenty-One Retrieval can assist all kinds of knowledge workers with their information needs. Examples of applications are electronic publishing (through the WWW), document information management, and workflow management. Possible user groups are, among others, publishers, researchers, marketeers, and private information consumers. The bottom-line in the exploitation of the Twenty-One technology will be the demonstration of its problem-solving potential to the audience: with the Twenty-One technology people should be able to work better, or reduce the ‘‘time-to-market’’ for information products.

5. Summary and conclusions

In this paper, an overview has been given of the Twenty-One approach towards electronic publishing via the Internet and/or CD ROM. By not relying on poor quality indexing, but by carefully coupling state-of-the-art multimedia processing, human language technology and various search techniques, a much more adequate environment for the dissemination of information can be realized. The employed technology seems highly supportive for the needs of the target user groups, and it is to be expected that, as a result of its usability within the domain of sustainable development, it will fulfil a crucial role in triggering both publisher and end-users in this, and similar domains, to use it. Both in terms of quantity, and in terms of quality, Twenty-One will contribute to a higher level of information exchange.

References

[1] J.E. Conklin, Hypertext: an introduction and survey, Computer (1987) 17–41.

[2] D. Hull, G. Grefenstette, A dictionary-based approach to multilingual information retrieval, in: Proc. 19th ACM Conf. on Research and Development in Information Retrieval (SIGIR96), in press.

[3] D. Hiemstra, F.M.G. de Jong, W. Kraaij, A domain specific lexicon acquisition tool for cross-language information retrieval, in: L. Devroye, C. Chrisment (Eds.), Proc. RIAO'97 Montreal, 1997, pp. 217–232.

[4] M. Ishii, K. Otha, H. Saito, An efficient parser generator for natural language, in: Proc. 15th Int. Conf. on Computational Linguistics (COLING), 1994, pp. 417–425.

[5] W. Kraaij, D. Hiemstra, Baseline tests for cross language retrieval with the twenty-one system, in: D. Harman, E. Voorhees (Eds.), Proc. 6th Text REtrieval Conf. (TREC6), National Institute for Standards and Technology, in press.

[6] W. Kraaij, R. Pohlmann, Evaluation of a Dutch stemming algorithm, in: J. Rowley (Ed.), The New Review of Document and Text Management, vol. 1, Taylor Graham Publishing, London, 1995, pp. 25–43.

[7] W. Kraaij, R. Pohlmann, Viewing stemming as recall enhancement, in: H.P. Frei, D. Harman, P. Schauble, R. Wilkinson (Eds.), Proc. 19th ACM-SIGIR Conf. on Research and Development in Information Retrieval (SIGIR96), Zürich, 1996, pp. 40–48.

[8] W. Kraaij, Multilingual functionality in the twenty-one project, in: D. Hull, D. Oard (Eds.), AAAI Symp. on Cross-Language Text and Speech Retrieval, American Association for Artificial Intelligence, 1997.

[9] D. Oard, B.J. Dorr, A survey of multilingual text retrieval, Technical Report UMIACS-TR-96-19, University of Maryland, 1996.

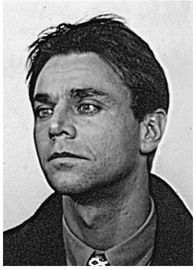
[10] R. Pohlmann, W. Kraaij, The effect of syntactic phrase indexing on retrieval performance for Dutch texts, in: L. Devroye, C. Chrisment (Eds.), Proc. RIAO'97, 1997, pp. 176–187.

[11] W.G. ter Stal, Automated interpretation of nominal compounds in a technical domain, Ph.D. Thesis, University of Twente, Enschede, 1996.

[12] M. Tomita, Efficient Parsing for Natural Language, Kluwer, Dordrecht, 1985.

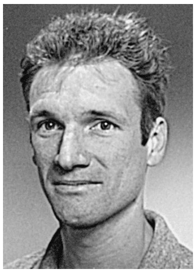
[13] M. Tomita, An efficient context-free parsing algorithm, for natural languages, in: Proc. IJCAI, vol. II, 1985, pp. 756–764.

[14] E.L. Trist, The sociotechnical perspective: the evolution of sociotechnical systems as a conceptual framework and as an action research program, in: A.H. Van De Ven, W.F. Joyce (Eds.), Perspectives on Organisation Design and Behavior, Wiley, New York, 1981, pp. 19–87.



Wilco ter Stal (1965) completed his M.Sc. in Linguistics and Literary studies, with specialisation Computational Linguistics in 1990. From 1990 until February 1996 he worked as NLP researcher at the Knowledge-Based Systems Group, University of Twente, Department of Computer Science, Enschede, the Netherlands. This period included a 3 months visit at the Microsoft Institute of Advanced Software Technology, Sydney, Australia. His PhD thesis entitled *Automated Interpretation of Nominal Compounds in a Technical Domain* completed the research period in Enschede. From May 1996 until now he is employed by Getronics Software, a major Dutch IT company, where he is involved in business development and advanced technologies. His interests include commercial speech and language engineering, knowledge management and innovative IT.

Franciska de Jong (1955) is professor of language technology at the Department of Computer Science, University of Twente, Enschede, The Netherlands. Furthermore she is supervisor of the research area Language, Speech and Information engineering of the CTIT (Centre for Telematics and Information Technology), Enschede, The Netherlands. She has a background in theoretical and computational linguistics. She worked with Philips Research as a researcher on the Rosetta machine translation project (1995–1992). She is advisor of several Dutch research institutes and has acted several times as reviewer for the European Commission.



Wessel Kraaij (1963) is a researcher of the Multimedia Group of TNO–TPD in Delft since 1995. His research interests are Language Technology and Information Retrieval and specifically Cross Language Information Retrieval. He's currently working on several European Multimedia projects including Twenty One and Pop-Eye. He graduated in 1988 at the University of Eindhoven (Institute of Perception Research) with a degree in Electronic Engineering (M.Sc.). Following his interest in HCI research he moved to the Institute for Language Technology and AI at Tilburg University where he stayed for 5 years. His main research focus there was the development of Natural Language Interfaces (ESPRIT II project PLUS). In 1994 and 1995 he worked at the research institute for Language and Speech (OTS/UIL) of Utrecht University where he developed and evaluated linguistic methods aimed at performance improvement of search engines. He's preparing a PhD on this topic.