w w w . e d e n t o o l . o r g

Summary of D6.2 Final Tool Validation and E-Democracy Evaluation Report

Project Number: IST 1999-20230

Project Acronym: EDEN

Title: Electronic Democracy European Network

Deliverable N°: D6.2 Summary

Version number: 1

Due date: N/A

Delivery Date: 23 July 2004

Short Description:

This report summarises evaluation of the pilots of software tools developed in the EDEN project. It complements the full report D6.2 *Final Tool Validation and E-Democracy Evaluation Report.* The pilots aimed to find out whether the tools would help to support communication between citizens and public administrations (PA's) in the cities of Antwerp, Bologna, and Bremen. The summary outlines evaluation criteria and targets, and the methods used to assess these. The results are discussed in terms of key questions that tie the criteria to the project's objectives.

Responsible partner: ITC (Napier)

Partners contributed: All

Authors: Angus Whyte, Ann Macintosh (ITC).

Made available to: CEC, Public.

Contents

1. Summary of Evaluation Results

1.1. Overview

This document is a summary of the report D6.2 *Final Tool Validation and E-Democracy Evaluation Report*, and complements that document. The EDEN project developed a range of tools intended to support communication between citizens and their public administrations (PAs), and better enable public participation in urban planning decision-making. The tools deploy Natural Language Processing (NLP) technology, to automatically process various kinds of text that are routinely used in those communications, namely:-

- Automatic routing to offices of citizens' messages according to their content ('*Address Guesser'*).
- Automated support for PA staff to manage Frequently Asked Question (FAQ) lists, for citizens to search them, and for their feedback to be used to update the FAQs ('*Answer Tree'*).
- Style checking tools for planning professionals to make documents easier for citizens to understand, by identifying "difficult" expressions and technical terms and suggesting alternatives from an organisation-wide glossary, including translations of terms into foreign languages for ethnic minority support. *('Style Enhancer*' and '*Multi-Language Helper'*).
- Natural language access to databases of maps and planning documents. ('*Natural Language Map'*)

These functions correspond to the EDEN tool named in brackets. Each can be applied to serve a purpose in its own right, but may optionally be integrated with a policy consultation 'front end', comprising:-

- Discussion fora with opinion polling and notification options to disseminate information according to a match between discussion themes and user profiles. ('*Guided Forum'*, '*Notification Handler*').

The pilot sites each had a desire to improve their online capabilities in the area of citizen engagement, but with preferences for different EDEN tools, and quite different deployment contexts. There is nothing in Natural Language Processing technology that makes it an inherently suitable tool for citizen engagement, and there are other measures that would help to accomplish that aim. The key assumption or working hypothesis of EDEN however has been that NLP can 'make a difference' when deployed for well-defined purposes as part of an infrastructure (human and technical) meant to support citizen engagement.

There are two aspects to the underlying logic of EDEN; firstly that NLP may *reduce the effort needed by citizens* to find relevant answers to their questions and understand them when they find them, and secondly that NLP may *reduce the administrations' effort* in handling the more routine communications involved in providing a response to citizens' concerns.

The validation of the tools and evaluation of the pilots was coordinated by Napier University's International Teledemocracy Centre. The other partners were involved as follows:-

Pilot Sites

- Antwerp (Digipolis): Validation of *Answer Tree*, and *Guided Forum*
- Bologna (Commune di Bologna) Evaluation of *Answer Tree* and *Address Guesser*
- Bologna (Archivio Osvaldo Piacentini) Evaluation of *Style Enhancer*
- Bremen (Planning Dept, TZI University of Bremen) Evaluation of *Answer Tree*, and *Guided Forum*
- Nisko (Infocentrum Nisko) conducted a feasibility study (included in the full D6.2 report) into the possibilities for deploying EDEN tools in Polish Public Administrations, but were not a pilot site.

Technical Partners

- Omega Generation: Support to pilot sites for deployment of *Address Guesser*, support for provision of evaluation data from log files, and comments on results;
- Public Voice Lab: Support to pilot sites for deployment of *Guided Forum;*
- *Yana Research:* Support to pilot sites for deployment of *Answer Tree, Multi-language Helper* and *Style Enhancer*, support for provision of evaluation data from log files, and comments on results.

The tools were piloted and evaluated using a framework of: -

- Evaluation criteria and indicators, developed with the participation of the PAs and consultation with their target users. The overall criteria were that the tools should support improvements to: -
    - Access in terms of the use of online information on urban planning, particularly by citizens who normally do not get involved in commenting on matters affecting their neighbourhoods.
    - Navigation and Comprehension, i.e. whether citizen-users manage to find and use documents (or office contacts details) that help them.
    - Acceptance, or the perceived legitimacy of the tools and their content as media for online participation in urban planning.
- Multiple methods to gather data on what citizen users did with the tools and what they thought about them. The methods were principally:-
    - Analysis of log files to compare the results citizens obtained with those the PAs expected, and apply standard measures for evaluating information retrieval performance.
    - Online questionnaires placed on the tool web sites to enable the test users to respond.
    - Interviews and questionnaires with users in the PA departments concerned.

Antwerp's Digipolis took part in *validation*, i.e. testing whether the tools were suitable for evaluation under real-life conditions – with the tools accessible on the city administration's web portals. However their pilots were curtailed, as Digipolis were not in a position to secure the commitment of the Antwerp PA departments concerned. Development of two of the tools was completed too late in the project for any of the sites to pilot them. They were *Natural Language Map* and the *Multi-language Helper* add-in module for Style Enhancer).

The table below shows the outcomes, focusing on the indicators for acceptance. It uses the following key: -

- ● Target that was fully met.
- ● Target that was not met, but results on the corresponding indicator were nevertheless considered good enough by the pilot site concerned.
- ❍ Target that was not met, and results were not thought good enough.
- ✓ Overall, the site judged the tool acceptable for possible further deployment.
- ✖ Overall, the site judged the tool unacceptable for possible further deployment.

The specific indicators that these targets refer to are given in the main section (2) of this summary, along with discussion of the results. More detailed indicators were also used for improvements to access, navigation and comprehension, which are also described and discussed.

Tool	*Pilot site*	*Targets*	*Overall outcome*
Address Guesser	Bologna	●●●●❍❍	✓ Thought promising though fine-tuning needed, and maintenance requirements need reduced.
Answer Tree	Bologna	●●●●❍❍	✓ Expected to reduce PA effort on direct handling of enquiries.
	Bremen	●	✓ Thought promising on basis of performance but insufficient time to gather user data.
Guided forum	Bremen	●●●●	✓ Positive in terms of citizen and PA acceptance of online consultation on neighbourhood planning, although tool maintenance effort too high.
Style Enhancer	Bologna/AOP	●	✖ Positive response from AOP participants but tool not precise enough for PA editorial requirements.

Table 1.1. Overall evaluation outcomes.

1.2. Sampling and user participation in the evaluation

The tools were evaluated between Dec '03- Jan '04 (except the Guided Forum pilots, which were carried out in two Bremen neighbourhoods from and 25/8 to 21/9 respectively). They were launched on the urban planning areas of the city websites, as prototypes intended for citizens whose need or interest was to communicate with the PAs during that period.

Citizen users took part on a self-selected basis, since the pilots aimed to test the tools in real-life conditions. The numbers of citizen users/ participants appear small relative to the populations of the cities, but not when considered against typical levels of enquiries/responses to the offices concerned. Bremen's pilot of Answer Tree was preceded by extensive validation using test queries, but suffered from server downtime, and was only available to citizens for 7 days.

Tool, Pilot site, Evaluation method	***Numbers of Participants***	
Address Guesser: Bologna	*Citizen users*	*PA users*
Log file of responses to 82 queries made by users.	70	-
Online questionnaire	70	-
Questionnaire & discussion: call centre & back office staff	-	4
Answer Tree: Bologna		
Log file of responses to 99 queries made by users.	45*	-
Online questionnaire	24	-
Questionnaire & discussion: call centre & back office staff	-	3
Answer Tree: Bremen		
Questionnaire & discussion: back-office staff	-	1
Usability field tests/ observations	2	-
Guided forum: Bremen (Horn-Lehe neighbourhood)		
Content analysis of 48 responses	70-100*	10
Online questionnaire	17	-
Questionnaire & discussion: PA officers	-	10
Questionnaire: Interest group representatives	-	5
Questionnaire: Elected representatives	-	15
Guided forum: Bremen (Waller Heerstrasse neighbourhood)		
Content analysis of 48 responses	35	-
Usability field tests/ observations	6	-
Online questionnaire	12	-
Questionnaire & discussion: PA officers	-	3
Style Enhancer: Bologna		
Comparison of 'difficult to understand' phrases in planning documents, identified by tool users & citizen readers.	12	6
Questionnaires & discussion: professional users	-	6

Table 1.2 Numbers of participants in evaluation

Table Notes:

- "-" means not applicable; " * " means estimated number of anonymous web users.
- Evaluation methods produced different kinds of data from the same individuals, i.e. samples overlap.

1.3. Summary of conclusions

To summarise the conclusions of the main report, the pilots showed that the NLP approach deployed in EDEN can reduce the barriers citizens and PAs face in communicating online, and in doing so may encourage those citizens who do not normally take part in city planning to contribute their views through online channels. In particular: -

1. The NLP approach implemented in *Answer Tree* was shown to be better at retrieving relevant FAQs in response to natural language queries than the widely used SWISH indexing and retrieval algorithm.
2. The *Address Guesser* tool showed promising results in finding relevant PA office addresses by comparing users' queries with those previously answered by them. Although the results were not accurate enough for users to be confident that the 'guessed' addresses were correct, refinements to the 'training' samples used and to the interface design appear likely to meet that objective.
3. The *Style Enhancer* tool was considered useful by planning professionals, particularly for checking relatively short documents giving general information to citizens. The tool is likely to be effective as a complement to a human editorial function although it is unlikely to replace that role. Further development of the glossary to differentiate between domain –specific and general usages of words and phrases would enhance the tool's effectiveness in that role.
4. The users who tested the tools, on a self-selected basis, were mostly satisfied with them and were mostly people who normally make enquiries by telephone or in-person. This indicated a potential uptake of online enquiry-handling estimated at 15% with wider deployment across other PA sectors than urban planning, although the short length of the pilots did not allow sufficient volumes of enquiry data to be used in making this estimate.
5. The pilot users mostly do not take part in city planning consultations by the traditional means (e.g. public meetings). Sizeable minorities of them agreed that the tools better prepared them to contribute their views online.
6. The *Guided Forum* pilots demonstrated that the recruitment of local citizens to help moderate online discussions can also help publicise neighbourhood or district level consultations to the citizens affected by planning decisions. The Bremen pilots also demonstrated acceptable levels of contribution quality and higher response rates than the traditional means, despite their small scale.

Recommendation for further research are proposed at the end of this summary report.

2. The Evaluation Approach

2.1. Introduction and Objectives of the EDEN Project

Governments are increasingly recognising a need to develop new methods to provide easier and wider access to government information and to achieve broader and deeper involvement of citizens in decision-making. The work of the OECD (Organisation for Economic Co-operation and Development) to promote frameworks for developing e-government and e-democracy (OECD, 2003), provides a useful framework that we will use to discuss the EDEN project outcomes, after first outlining the project aims.

The EDEN (Electronic Democracy European Network) project was funded through the European Commission's Fifth Framework Programme under the thematic programme 'Systems and Services for the Citizen' which specifically includes R&D projects aimed at e-democracy. EDEN was a collaborative project with public administrations (PAs): Bologna, Antwerp, Bremen, Nisko, and Vienna, along with the Bologna based Osvaldo Piacentini Archive, and with research partners: Omega Generation, International Teledemocracy Centre, Public Voice Lab - PVL, Digipolis Antwerp, TZI - Centre for Computing Technology at the University of Bremen, and Yana Research.

The overarching objective of the EDEN project was to stimulate and support citizens' participation in the decision-making process, specifically in the area of urban planning. The project objectives do not explicitly define ' improved participation'. They do so implicitly in terms of enabling more and 'better informed' questions, comments, requests, or complaints to be made online, by citizens who may be affected by plans but do not normally get involved. This 'e-enabling' should be achieved by providing information that is more accessible and easier to comprehend and navigate.

EDEN focused on urban planning partly because it is an area of public administration that has a longer history of citizen participation than most. Planning is also a useful test domain for e-government because the policies and procedures directly involve private citizens and businesses, and define how and when others may have a say in those decisions. Thus EDEN overlapped the public and private spheres, and its software tools needed to address the communication needs of both.

Urban planning was the focus of the pilots, the results of which may be applied in other administrative domains. That is, the tools being developed are not urban planning applications but are intended for broader use. The pilots each involved deployment of EDEN tools in various combinations on the websites of the city administrations, to allow them to be used on an 'experimental' basis by citizens, public officials and planning professionals.

This report summarises the conclusions drawn by the main authors of the report Napier University's International Teledemocracy Centre, after consulting the other partners in the project who carried out the pilots of EDEN tools in their respective cities, provided evaluation data and an account of their experience, and commented on the outcomes.

A key assumption or working hypothesis of the project is that *Natural Language Processing* (NLP) tools, when integrated into the PA's (Public Administration) infrastructure for communicating with citizens on matters related to urban planning, can 'improve' public participation in decision-making. Thus it is important to note that the project not only began with the premise that NLP technology could address the objectives, but that to do so it would need to be applied in the form of tools with a specific set of functions:-

- Automatic routing to offices of citizens' messages according to their content ('*Address Guesser'*).
- Automated support for PA staff to manage Frequently Asked Question (FAQ) lists, for citizens to search them, and for their feedback to be used to update the FAQs ('*Answer Tree'*).
- Style checking tools for planning professionals to make documents easier for citizens to understand, by identifying "difficult" expressions and technical terms and suggesting alternatives from an

organisation-wide glossary, including translations of terms into foreign languages for ethnic minority support. *('Style Enhancer*' and '*Multi-Language Helper'*).

- Natural language access to databases containing maps and planning documents relating to the city neighbourhoods. ('*Natural Language Map'*)

These functions correspond to an EDEN tool with the working name in brackets. Each can be applied to serve a purpose in its own right, but may optionally be integrated with a policy consultation 'front end', comprising:-

- Discussion fora with opinion polling and notification options to disseminate information according to a match between discussion themes and user profiles. ('*Guided Fora & Polling'*, '*Notification Handler*').

The tools were piloted for approx. one month (December '03- January '04), and preceded by validation tests. Interim results from these validation tests were reported in EDEN Deliverable D6.1, and updated in D6.2. The validation of the tools and evaluation of the pilots was coordinated by Napier University ITC, and the roles of the other partners were mainly as follows:-

Pilot Sites

- Antwerp (Digipolis) Pilots of *Answer Tree*, *Guided Forum* and *Natural Language Map*
- Bologna (Commune di Bologna) *Answer Tree* and *Address Guesser Pilots*
- Bologna (Archivio Osvaldo Piacentini) *Style Enhancer* and *Multi-language Helper* pilot
- Bremen (Planning Dept, TZI University of Bremen) Pilots of *Answer Tree*, *Guided Forum* and *Natural Language Map*

Technical Partners

- Omega Generation: Support to pilot sites for deployment of *Address Guesser*, support for provision of evaluation data from log files, and comments on results;
- Public Voice Lab: Support to pilot sites for deployment of *Guided Forum;*
- *Yana Research:* Support to pilot sites for deployment of *Answer Tree, Multi-language Helper* and *Style Enhancer*, support for provision of evaluation data from log files, and comments on results.

The pilots are described in chapters 3-6 of the D6.2 report, according to the framework developed in collaboration with the project partners. It is perhaps surprising that published evaluation frameworks and studies of e-democracy impact are notable for their rarity. The evaluation task is often hindered by lack of clarity in objectives, lack of definitions and indicators of success, the complexity of the relationships between stakeholders, and barriers to reporting both failure and success (OECD, 2003). Yet there is widespread agreement of the need for sound evaluation, given the potential impacts of ICT in alignment with organisational change. As Fountain remarks, those impacts raise fundamental and important questions for central concepts of governance such as accountability, task specialization, and jurisdiction (Fountain, 2002).

EDEN tools potentially impact on administrative functions and services, particularly the handling of enquiries, that extend beyond the use of discussion fora that are commonly associated with the term e-democracy. In the conclusions at the end of this chapter we discuss how the e-government (service related) and e-democracy (citizen engagement) aspects of EDEN are inter-related, through changes in role for the stakeholders involved, and through issues of how citizens represent themselves. We also discuss EDEN's outcomes in terms of a trajectory from information provision through consultation to participation, a central theme of the OECD's framework that we turn to next.

2.2. A Trajectory from Information Provision to e-Participation

The OECD define three types of interaction associated with citizen engagement (OECD, 2001) that have become widely referred to, namely:

> **Information**: a one-way relationship in which government produces and delivers information for use by citizens. It covers both "passive" access to information upon demand and delivers information for use by citizens and "active" measures by government to disseminate information to citizens.
>
> **Consultation**: a two-way relationship in which citizens provide feedback to government. It is based on the prior definition of information. Governments define the issues for consultation, set the questions and manage the process, while citizens are invited to contribute their views and opinions.
>
> **Active participation:** a relationship based on partnership with government in which citizens actively engage in defining the process and content of policy-making. It acknowledges equal standing for citizens in setting the agenda, proposing policy options and shaping the policy dialogue – although the responsibility for the final decision or policy formulation rests with government.

These distinctions indicate a scale of 'engagement' in policy-making along which government initiatives could be plotted. That is how its authors use it, in reporting that "efforts to engage citizens in policy-making on a partnership basis are rare, undertaken on a pilot basis only and confined to a very few OECD countries" (ibid.). Applying this principle to the OECD definitions above, Macintosh (2004) describes three levels of participation that can be used to characterise e-democracy initiatives: -

> **E-enabling** is about supporting those who would not typically access the internet and take advantage of the large amount of information available. The objectives we are concerned with are how technology can be used to reach the wider audience by providing a range of technologies to cater for the diverse technical and communicative skills of citizens. The technology also needs to provide relevant information in a format that is both more accessible and more understandable. These two aspects of accessibility and understandability of information are addressed by e-enabling.
>
> **E-engaging** with citizens is concerned with consulting a wider audience to enable deeper contributions and support deliberative debate on policy issues. The use of the term 'to engage' in this context refers to the top-down consultation of citizens by government or parliament.
>
> **E-empowering** citizens is concerned with supporting active participation and facilitating bottom-up ideas to influence the political agenda. The previous top-down perspectives of democracy are characterized in terms of user access to information and reaction to government led initiatives. From the bottom-up perspective, citizens are producers rather than just consumers of policy (Macintosh et.al, 2002). Here there is recognition that there is a need to allow citizens to influence and participate in policy formulation.

These terms are helpful since the term 'participation' can be applied to *any* of the levels of engagement- to refer to the extent that citizens make active use of the information that PAs intend to be 'e-enabling', or in relation to their use of e-engagement or e-empowering tools (such as online citizens juries, or online petitioning). On this scale EDEN's city administrations *declared objectives* for the tools relate to 'e-enabling' and 'e-engagement'.

2.3. Aims of the EDEN Tools

Three of the five main tools, *Answer Tree* and *Address Guesser* and *Natural Language Map* entail the *processing* of citizens' enquiries by computational linguistics technology, or Natural Language Processing (NLP). The *Guided Fora* tool does not use NLP directly but takes the more conventional form of 'threaded conversation' found in discussion fora. Each of these tools was intended for any user of the city administrations' web sites, and they were deployed for that purpose in the pilots reported here. NLP is also

used in the *Style Enhancer* writers' tool, and its translation aid *Multi-Language Helper*. These were intended to be used online by PA users or contracted planning professionals.

In this section we focus on the options for communication that the tools offer when integrated into the web sites of the cities involved, and the benefits sought. These are depicted in Figure 2.1 below.

EDEN envisages that citizens with a question, comment or complaint (etc.) can choose to go online and send a private message to the administration via *Address Guesser*, perhaps first checking to see whether their concern is already addressed by a FAQ (Frequently Asked Question) in *Answer Tree*. As the name suggests, this allows a topic tree to be browsed to find an answer. It can also be searched however, and *Answer Tree* is meant to respond effectively to questions in natural language (as in Figure 2.1) without the user needing to know what keywords to use, or use Boolean operators to combine them for best results. Similarly, if they want or need to find background information relating to a particular geographic part of the city they may enter a natural language query in *Natural Language Map*. This identifies keywords and performs a search on a database repository, which in turn relates documents to the spatial coordinates of maps held in the administrations' GIS (Geographic Information System). Matching results are displayed in the form of a map with links to related documents.

If a suitable answer is not there, users may email the Administrator of the *Answer Tree* to suggest that it be added, and receive a reply – although probably after the message has been routed to the office specialising in the matter concerned. If the citizen already knows which office that is they may send an email directly to it instead of via the *Answer Tree* Administrator. However many citizens neither know nor care about the administrative structures. In that case *Address Guesser* provides suggestions of the correct office(s), based on a comparison between the content of the message, and that of a compiled set of previous emails answered by each office. The citizen may however prefer to know what other citizens think. The *Guided Forum* tool provides threaded discussion, typically focused on district or neighbourhood-level topics (which may be added to on users' suggestions), and supported by related planning documents.

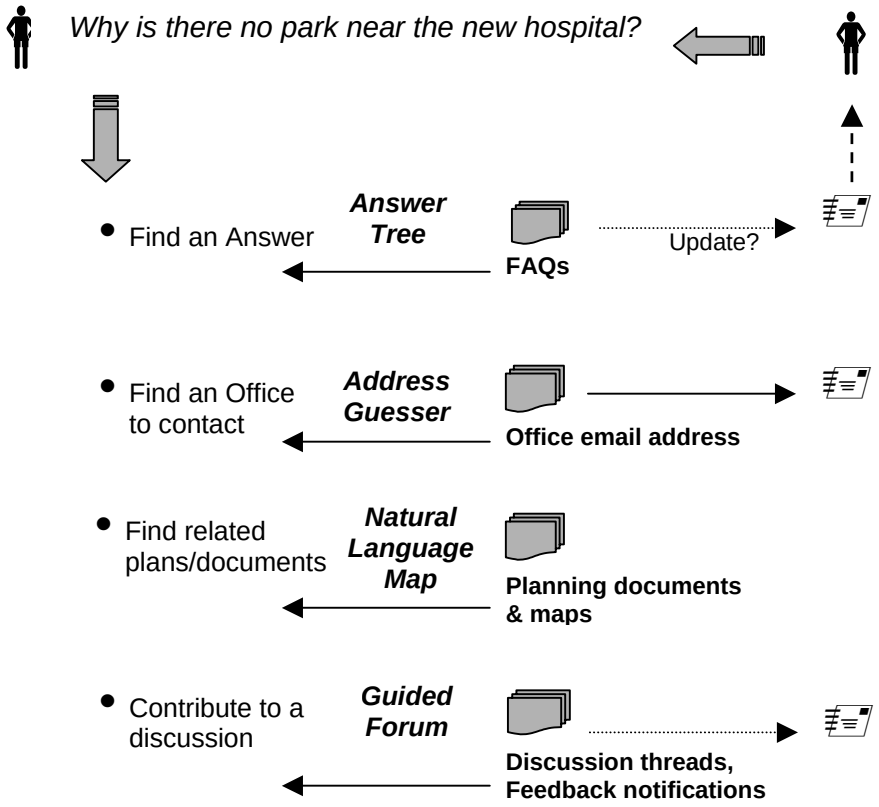

Figure 2.1. Options for communication between citizen and administrations

EDEN thus inter-relates the handling of *private* enquiries, a service typically associated with e-government, with the more *public* discussion normally associated with e-democracy. Benefits are sought for both parties to the communication- citizens (as individuals or civic groups) and for the 'back office' functions and officers, whose role is to respond or proactively seek citizens' feedback. These hoped-for benefits can be summarised as follows:-

Style Enhancer is not shown in Figure 2.1 as citizens are not direct users. Rather it was used by officials and planning professionals (e.g. architects) who produce documents for a target audience that includes 'ordinary citizens'. It should help them improve accessibility of those documents by identifying how grammatically complex phrases can be replaced with simpler wording, and identifying technical jargon that can be replaced with simpler alternatives retrieved from a glossary. *Multi-Language Helper* is an 'add on' to this tool, which retrieves translated phrases from the glossary and, on demand, displays the translations wherever the phrases appear in an online document. The benefits expected from deployment of the tool are that more people can easily understand the documents that are published (for *Multi-Language Helper*, referring to speakers of the target languages included in the online glossary), and that officers are more confident that difficult expressions in planning documents can be understood by citizens who usually find them difficult.

Address Guesser should support enquirers by providing faster and more accurate targeting of their enquiry when they do not know whom to contact, thereby reducing the PA effort on 'front office' enquiry forwarding functions.

Answer Tree should stimulate an increase in the number of people accessing online information on decision-making procedures and outcomes. Its use should allow answers to 'frequently asked' questions about planning to be more easily published on the City website. This should result in more people finding relevant answers to their questions on the website, and more people making better informed choices on whether to contribute their own views about planning matters.

Guided Forum. The PAs aim to increase citizens' engagement in planning, leading to avoidance of (long-term) planning mistakes arising from inadequate participation of those affected by decisions. Online consultations, i.e. discussion fora that are focused on particular planning decisions and which guide citizens through the relevant decision-making phases and background information, should support this mainly by providing additional channels for public discussion between the PA and the citizens, enhancing the capability to involve those legally entitled to a view.

The *Notification Handler* option for *Guided Fora* is intended to improve the efficiency of communications between citizens and administrations about the consultation process. Firstly, by alerting citizens to forum events that match interests they have specified. Secondly, by providing administrations with a cost-effective means to inform particular geographic and interest-based target groups about planning decisions that may affect or interest them.

2.4. Outline of the Methodology

The EDEN tool evaluation was time-limited and it is more appropriate to think in terms of probable benefits and risks rather than impacts. In this section we describe the methodology used to tie the needs and benefits outlined in the previous section to more specific targets, and to risks representing barriers to the desired impact.

2.4.1. General Approach

The evaluation framework rests on several methodological assumptions that we should make clear. One of the more fundamental is our preference for a case study approach rather than a statistical experiment. We discuss the reasons for this elsewhere (Whyte and Macintosh, 2002), but briefly the social research in EDEN is exploratory, e-democracy evaluation methods are in their infancy, and there are few previous evaluations that can attest to the validity or reliability of relationships between quantifiable variables. That is not to say that we made no use of quantitative data – many of the measures we used are quantified. However with the

EDEN tools deployed on a pilot basis, it was neither desirable nor feasible to try to control the factors that influence their use. Our aim was rather to add to understanding of those factors, and to use case study strategies for maximising validity, namely the use of multiple sources and methods, and a traceable path from evidence to conclusions (e.g. Yin, 1994).

Our approach began by combining action research, ethnography and evaluation approaches, a hybrid based on Suchman and Trigg (1991). Our *action research* role was initiated in the early requirements gathering phase of the project, when we worked with project members in each city to elaborate on and clarify the rationale for change, and to coordinate their work with citizens to understand the requirements for the tools. This work drew on Soft Systems Methodology (SSM) (Checkland and Scholes, 1990) and scenario-based methods (Carroll and Rosson, 1992;). These are not discussed in detail here but involved structured discussion with PA officers and citizens about current online activities and the proposed changes to them, i.e. we related the benefits and consequences that our scenarios claimed EDEN tools could have, to how the scenario readers supported or criticised them, or offered alternative proposals. In parallel with the discussions this involved, we surveyed citizens about communications issues and their interests in EDEN. These materials provided a broad understanding of what was thought feasible and desirable about the design proposals, and why. The results of that work were described in the project deliverable *D2.2 User Requirements Analysis*.

The role of *ethnographic* methods in the project has been relatively small, but important to understanding how communications are currently accomplished. In systems design contexts ethnography normally involves observations, semi-structured interviewing, and detailed analysis of recorded interactions between the actors involved in accomplishing work in its everyday setting (see e.g. Suchman and Trigg op.cit.). In our case we relied on samples of emails and contributions to discussion fora, alongside the accounts given by planning professionals and communications officers in semi-structured interviews about their work. These informed the system requirements, together with the evaluation indicators and the deployment risks and issues that we used to focus the evaluation questions. The results of that work were described in *D6.1 Interim Tool Validation with Refined Requirements Specification*.

2.4.2. Evaluation Questions, Criteria and Methods

The main evaluation questions then were:-

1) Did the NLP-based information retrieval tools provide relevant answers to citizens' questions?
2) Did the pilots demonstrate the anticipated improvements to online access, navigation, and comprehension?
3) To what extent were the tools accepted by citizen and PA users, and why?
4) Did the tools better enable citizens to contribute views on their neighbourhood or city planning?

These questions address whether or not EDEN meets its aims, and for the last question in particular there are important dimensions that need addressed: -

- What comparisons can be made between the traditional offline methods and online methods with EDEN?
- Were the skills & capacity available to use the tools effectively, from citizens' and PA perspectives?
- For the PA, what were the conditions that needed to be met for full deployment?

2.5. Evaluating whether NLP tools give relevant results

2.5.1. Measures for evaluating retrieval

Answering our first evaluation question began from established practice in assessing the performance of information retrieval systems. Evaluation of the capabilities of such systems to give results that are relevant to what the user is looking for – expressed as a query –is exemplified by the work of the Text Retrieval Conference (TREC), some of which assesses best practice in applying the standard evaluation measures. These are *precision* (the number of relevant documents in the results set, divided by the total number of documents retrieved), and *recall* (the number of relevant documents in the results set, divided by the total number of relevant answers present in the system) (e.g. Voorhees and Harman, 1999).

The values of precision and recall are inversely related, that is there is usually a trade-off between high scores on one and the other. The commonly accepted way of representing the results of this trade-off are to graph their respective values at successive 'cut-off points', in other words after (e.g.) 2, 4, 6, 8, or 10 results have been assessed for each query, i.e. a *precision-recall graph*. A summary measure, *average precision*, is often used to represent the area under a precision-recall graph (Buckley and Voorhees, 2000).

When evaluating performance the accepted practice is to compile a *test collection* comprising a set of queries that represent a range of topics, a set of documents on those topics, and a set of judgements about which of them are relevant to each of the queries. Test collections are the basis for 'laboratory based' (i.e. controlled) experiments- for example to carry out several test runs using several alternative search methods , while keeping unchanged the test queries, documents, and relevance judgements (ibid.).

Recent research includes the evaluation of *question-answering* systems; those specifically intended to retrieve a limited set of answers to questions, rather than long lists of search 'hits' that have varying degrees of relevance to a topic. In that respect both *Answer Tree* and *Address Guesser* can be considered as question-answering systems. In *Answer Tree*, the user is presented with a hierarchically organised list of 'frequently asked' question, each with a short document intended to answer it. Instead of browsing, the user may want to search for a question that most closely matches his or her own. With *Address Guesser* on the other hand, the user's typical question will of the form "which office can answer my question".

A variant of the precision measure used to evaluate question-answering systems is the *reciprocal rank score* (Radev et al, 2002). This is designed to take account of the position of relevant answers in the results set, i.e. it gives more weight to a results set that has a relevant answer shown first than one where only (for example) the second result is relevant. The reciprocal rank score is the sum of the reciprocal ranks of all the relevant responses. For example, if a query gives a set of 3 question-answer pairs, and only the second and third are relevant, the score is 0 for the first, 1/2 (0.5) for the second, and 1/3 (0.33) for the third, giving a reciprocal rank score of 0.83. Normally the tests for reciprocal rank (as for average precision) are carried out for 50-100 test questions, and a mean score obtained.

2.5.2. Applying the Measures for 'Fine Tuning' and Validation Purposes

We identified the Mean Reciprocal Rank Score as an appropriate summary measure for the 'question answering' tools *Address Guesse*r and *Answer Tree,* and defined targets for mean precision and mean recall. We revised the measures to include Mean Average Precision since this is a more generally applicable measure that corresponds to the precision-recall graphs recommended by the project's Review Panel. These measures were also intended to be used with *Natural Language Map*, but this unfortunately could not be piloted in time for reasons discussed later. The measures are not so readily applied to other tools, although recall and precision can in principle be applied to *Style Enhancer* if the 'documents to be checked' are treated as queries to a system that retrieves advice messages (an approach that we considered but rejected in favour of simpler user ratings).

These measures were applied to the EDEN tools first of all, as an 'internal' quality check to ensure that the data (FAQs for *Answer Tree*, office email samples for *Address Guesser*) had been properly set up to get the

best results from the NLP 'linguistic resources'. These resources include glossaries covering urban planning terms, and software representations of grammatical rules for Italian, German and Dutch/Flemish. The software automatically 'parses', i.e. applies those rules to, whatever text is entered by the user. It does this in order to distinguish the relevant terms and syntactic structures (e.g. 'noun phrases') that identify a matching *Answer Tree* FAQ, or if using *Address Guesser*, an office that matches previous examples of email enquiries. The linguistic resources have to be 'tuned' to work at their best and the validation tests helped to quantify the success of the tuning. The tuning tasks included changes to the grammatical rules and vocabulary, and defining a suitable range of 'stop words', i.e. common words that would not help differentiate relevant from non-relevant results. Another reason for testing was to assess the difference that *synonym handling* made to *Answer Tree* performance.

Finally and most importantly, the validation tests allowed an assessment of the *added value of the NLP parser*. As we mentioned already, *Answer Tree* first uses the NLP parser to break down what the user types, into the appropriate terms. Then those terms are entered (unseen by the user) into a search on the FAQs, using the SWISH algorithm. SWISH (Simple Web Indexing System for Humans) is 'open source' text indexing and retrieval software that is commonly used in web search engines. It is this software that actually returns the search 'hits' in *Answer Tree*. Any added value from using NLP therefore comes from its ability to first select the 'best' search terms from a question expressed in 'natural language', rather than the user having to choose keywords to express what he/she wants to look for and combine them with Boolean operators (e.g. "(motorbike NEAR restrictions) AND NOT bicycle"). To allow that value to be tested, the technical partners provided "no parser" versions of *Answer Tree* in German and Italian.

To perform the validation test runs, each pilot site:-

1. Wrote 100 queries to cover the topics of their *Answer Tree* FAQs, and/or offices involved in testing *Address Guesser*.
2. Identified the expected results (FAQs or offices) for each query, listing them in a spreadsheet.
3. Fed each test query into their software, for each test run.
4. Napier then retrieved the results, matched them against the expected results in the spreadsheet, and calculated the results of the measures discussed above for each query and the test run average.

The results of these tests are shown alongside the evaluation results in section 2.8.1

2.5.3. Applying the Measures for Evaluation Purposes

The key difference between the evaluation and validation tests was that while the first task of validation was to write the test queries, the evaluation was carried out *using queries entered by citizens themselves*, and extracted from the log files of the *Answer Tree* and *Address Guesser*. These held (anonymously) the text entered by every user of the EDEN tools during the pilot periods when they were online on the cities' web sites. The log files also listed the FAQs or office addresses that the citizen/user's got in response to their queries.

Using the citizens' queries was a departure from normal practice, but a key element of the evaluation since it allowed us to compare 'natural language' typed by citizens with the conception of 'natural language' formalised in the NLP software. As we pointed out in the previous section, test queries are normally written by specialists to ensure that the breadth of topics is evenly covered, and that the queries test the capabilities that are being validated (e.g. spelling errors are not included unless spell-checking is being tested). Since the NLP offered the capability to match queries with relevant results that used different wording, the Answer Tree test queries were worded so they did not exactly match the wording used in the FAQ itself. A common feature of both however was that the queries and the results were syntactically 'well-formed'.

The idea of 'well-formed' syntax is an important one in linguistics and although we do not have the scope to discuss it in detail, it corresponds to the idea underpinning NLP; 'natural language' is that which follows known grammatical rules for a language. As discussed in a separate report (D4.2, Application of NLP tools to the field of e-democracy services) this became a contentious point in the *validation* testing, since several

pilot sites wanted to use texts that had been written by citizens regardless of whether they were grammatical or not, on the grounds that the software should process language of the kind actually used by citizens.

Since the NLP software can only process 'well-formed' texts, to include 'badly-formed' texts in the validation test collection would lead to poor performance without testing the capabilities the NLP was meant to have. The greater appreciation of that fact among the pilot sites also however led to more realistic expectations of the tools' performance when faced with citizens' actual language.

The precision (or reciprocal rank) and recall measures are not so readily assessed for *Natural Language Map* and *Style Enhancer/ ML Helper*. This is because the measures depend on a judgement about what the 'relevant results' should be. Users will each make their own judgement about that, so by definition the judgements cannot be applied consistently across the queries used for the 'test run'. By contrast, the relevance judgements can be made relatively objectively for *Answer Tree* and *Address Guesse*r since it is more likely that the FAQs or office addresses that are the appropriate answers to a query can be identified by the PA, without the need to know more about what the user is trying to achieve.

For the *Style Enhancer/ ML Helper* 'writers' aid' tools the issues were: -

- To what extent were the messages displayed understandable and helpful to the users?
- To what extent did professional authors/editors and citizen readers of urban planning documents find the same phrases problematic in any given document?
- To what extent did that matter? i.e. Did both author/editors and citizens find the documents easier to understand after 'style enhancing'.

These are questions that we ask under the wider criteria of better comprehension and overall acceptability described next.

2.6. Changes in Access, Navigation, Comprehension and Acceptance

In this section we outline criteria, indicators or targets, and sources of data that we developed to address the broader question of system acceptability (cf. Nielsen, 1990). These are listed at the end of this chapter where we discuss the outcome of their assessment. So far we have considered evaluation from the technical perspective of information retrieval research. This is clearly not enough, but if we follow the idea of a trajectory from better information provision to better acceptance of online tools in the decision-making process, we can see better information retrieval performance as a "necessary but insufficient" condition for meeting the aims of EDEN. Whether or not an information retrieval system achieves high precision and recall in relation to the test of a query, what users are interested in is whether or not the results help them meet their overall information-seeking needs (Spink, 2002). In other words while retrieval evaluation assesses performance in relation to *queries*, we are concerned with how citizens' *enquiries* are handled and the two are not necessarily the same.

The EDEN tools are intended, as we have said, to enhance *access*, *navigation* and *comprehension*. These can be considered as conditions for meeting the larger goal of increased acceptance of online tools by citizens wanting a say in decision-making, and by policy-makers responsible for those decisions. These criteria need more exact working definitions, and to arrive at them criteria and targets were drafted and discussed among the city partners, and with citizens. As a result, indicators and targets were developed based on the following basic definitions:-

> *Access* concerns how much information was made available, who accessed it, when and how often. When assessing this, we emphasised access by people who say they normally do not get involved in commenting on matters affecting their neighbourhoods.
>
> *Navigation* and *Comprehension* concern active use of the EDEN tools after access is made. That is, once users gained access, did they manage to find documents that helped, or offices to contact?

And what did they make of what they found, in terms of being able to comprehend and act on what is said?

Acceptance concerns the perceived *legitimacy* of the tools and their content as media for online participation in urban planning.

The indicators and targets were broken down from each of these criteria and related to each of the tools. They are outlined later in this summary. The main sources for the evaluation were:-

- Samples of messages between citizens and officers; contributions to discussion fora, and other operational data resulting from use of the tools.
- Interviews with officers and citizens invited to 'user panel' group interviews, and individually.
- Usability tests in the field, through observing users trying the tools and recording usability problems and their severity.
- Questionnaires: both in print form, with user panel participants, and online to allow any user to respond.
- Log files: the tools log all queries to *Answer Tree* and *Address Guesser*, and the responses provided by the system (FAQs and office addresses, respectively). Examples of queries and results were analysed qualitatively as well as quantitatively.
- Web server log files also provided details of page requests and visits to each tool, indicating for example whether *Guided Forum* users also sought information about the forum topics.

These sources were also used before the evaluation to establish the requirements and identify what the indicators should seek to establish, and the trade-offs between them. For example *Answer Tree* provides a route to a general enquiry handling office, the "Administrator", as shown in Figure 1. A key design assumption is that the messages that citizens send are not those that the user could find by browsing the FAQ 'tree', but instead are non-routine questions that may be useful for new FAQs, and alleviate the enquiry-handling function of the more routine ones. But if the search function has low performance, or if users cannot easily browse to find an answer, there is a risk the Administrator could be inundated – in effect a risk of success on the 'access' criterion at the expense of failure on 'acceptability'.

For the *Guided Forum* tool, the criteria are consistent with the interest of the Bremen city partners who piloted it, in promoting *deliberative* discussion of local plans and related issues. Concepts of deliberative democracy underlie much of the benefits sought for e-democracy, and in EDEN's case are translated into a rating scale used by researchers and public servants to assess the contributions that citizens make online, as follows below (Westholm, 2003). These ratings helped the administration's assessment of whether the forum outcomes were acceptable.

2.7. Evaluating the Probable Impact on Citizen Participation

2.7.1. The Questions Addressed

In our evaluation of the EDEN tools we are not only interested in the satisfaction of individual pilot site users with their visit to the corresponding web site, whether those users are in the administrations involved, the 'ordinary citizens', or other anticipated user groups such as architects working on public projects. Our interest is also (or more specifically) in these questions: -

- Does the experience of using the tool lead citizen users to feel either *more involved* in local policy making, and/or better prepared to contribute their views by being *better informed* ?
- Does the experience of being involved in the pilot lead PA users to feel *better prepared and resourced* for citizen engagement in city planning decisions?
- Do citizens or other users see other (met or unmet) benefits in participating online?

- Are the users who have a positive view of EDEN's impact those whom the project targeted, and if not how do they differ?

These are assessed partly from the results of questionnaires, including 'exit questionnaires' placed online alongside the EDEN tools, partly from interviews with users, and partly from the participant-observation of members of the pilot site project teams.

2.7.2. Whose Impact? The Sampling Approach

An important issue in the evaluation is to what extent the views obtained are representative of the groups that EDEN has sought to involve. This is also a methodological issue since the nature of the 'sample' of citizen and other users involved in the project needs to be considered alongside the validity of the findings.

In general terms EDEN targets people who work, live or have some other interest in the pilot cities and more especially those neighbourhoods affected by plans that were put to public consultation within the project's lifetime. The project objectives claimed that by simplifying access to information EDEN would reach the 'silent majority generally left out of PA processes'. So EDEN is particularly interested in attracting people who are already Internet users but do not normally use it to take part in decision making, and also do not normally 'get involved' in the traditional ways, such as by attending public meetings or writing directly to an elected representative.

The sampling approach took into account substantial d*ifferences between the pilot sites* and the exploratory nature of the research. So for practical reasons as well as methodological preference the study has used *purposive* sampling. That is, the test users have been selected (and self-selected in the case of online questionnaires) by purposefully seeking whichever citizens the PA needs to communicate with (or who need to communicate with the PA) at the time of the pilot. In effect that has meant that questionnaire respondents have been self-selected, and the response rates heavily dependent on the availability to the Public Administrations of *contactable* citizens, interested enough in the possible benefits of the tools to visit the relevant home pages (linked to from the PAs main site or portal), try out the tools, and then complete a questionnaire. We return to the sampling issues in Section 2.10.3 'Methodological strengths and weaknesses'.

2.8. Differences between the Pilot Sites

The Public Administrations involved in EDEN identified different needs and priorities for the various software tools, and applied them for different contexts. These differences were expected, alongside some differences in the methods that could be applied. There were also unexpected and unplanned differences, which limited the data sources available and in some cases prevented pilots being carried out. Below we give a brief overview of each pilot site focusing on its similarities and differences from other pilot sites, and the background to its selection of tools, while each city's pilot report gives a fuller account in Chapter 4 onwards.

2.8.1. Bologna. Commune di Bologna (CoBo).

Bologna municipality was represented in the project by the Office for Relations with the Citizen (URP). URP responsibilities include managing the information and services provided by the Iperbole Civic Network, handling enquiries at the 'front office' desk and via telephone at the Call Centre.

The Iperbole Network is unique in EDEN since the Municipality has since 1995 provided free Internet services (for example email accounts) that in other participating cities would be provided by numerous 3rd party service providers. The base of approx. 18.000 subscribers gave the project opportunities to recruit citizens that were not available to other pilot sites.

Unlike the other participating PAs CoBo has no single department or office responsible for urban planning matters. These responsibilities are distributed between 13 sectors of the municipality. Also although other pilot sites have a central Call Centre function, it was only in Bologna that they participated in the evaluation.

Address Guesser was piloted only in CoBo, which for the reasons given above has a suitably sized test domain in terms of the number of offices involved in urban planning. CoBo also has an existing 'email routing' application CSS (Customer Satisfaction Service) that provides a baseline for comparison. Quantitative comparisons of performance cannot be made however since CSS is not deployed in most of the offices related to urban planning. Qualitative comparisons are also difficult since CSS works by matching email texts with keywords that are manually recorded against the various office descriptions, rather than through any automatic analysis to identify keywords.

Answer Tree was piloted for various planning-related topics, although unlike in Bremen and Antwerp these were not focused on particular urban planning developments in specific neighbourhoods. Rather the focus was on topics of more general interest, both to planning professionals and 'ordinary citizens'.

Style Enhancer and to a more limited extent *Multi Language Helper* were piloted with the assistance of Archivio Osvaldo Piacentini (see below).

2.8.2. Bologna: Archivio Osvaldo Piacentini (AOP).

The AOP is a non-profit association that works in the Reggio Emilia region around Bologna in the fields of urban planning, public administration, cultural promotion and research. In EDEN it provided a 'gateway' to planning professionals (architects in particular) with an interest in the outcomes of the pilots. AOP was not in itself a pilot site, rather its role was to assist CoBo in pilots of the *Style Enhancer* and *Multi Language Helper*, including the set-up of a glossary of planning terms, one of the electronic resources used by both tools to assist authors/editors in making documents more comprehensible to a lay audience.

2.8.3. Bremen: Planning Dept and University of Bremen TZI

The Department for Urban Planning and Building Regulations is part of the Senate for Building and Environment in Bremen, whose remit covers all topics from urban planning concepts and citizen participation to detail planning and building permits. In EDEN the department worked closely with TZI, a research centre at the University of Bremen. TZI is an inter-departmental centre focusing on application-oriented research and development in computer science.

Answer Tree was selected for piloting in Bremen following promising initial tests from Antwerp. Late in the project Bremen decided to test whether the tool could enhance the provision of information about its neighbourhood consultation, for users of the Guided Forum.

The *Guided Forum* was piloted in two different neighbourhoods, Horn-Lehe and Waller Heerstrasse, to support consultation on future developments in those neighbourhoods.

2.8.4. Antwerp Digipolis

Digipolis was formed in 2003 from the merger of two (non profit) organisations providing ICT support services to the Public Administrations of Antwerp and Ghent. It's role is unlike that of other pilot sites in that the City of Antwerp is not a direct participant in EDEN. Digipolis' role in the pilots therefore depended on the support of one or more sponsoring departments in the PA.

The *Answer Tree* was deployed on a web site to provide FAQs about the reconstruction of *De Leien* a major thoroughfare and traffic route through the centre of Antwerp.

The *Guided Forum* was deployed to support changes in the PA's procedures for Neighbourhood Consultation. The Forum was intended to be deployed for about 15 'Urban Neighbourhood Consultation' neighbourhoods that are in phase 3 - "execution" of the consultation cycle from mid-July 2003. This was intended as a channel for citizens' comments on ongoing projects rather than for influencing those decisions that have already been taken before the execution phase.

These pilots were unfortunately curtailed. We understand that this was because of difficulties Digipolis faced in securing the commitment of the PA departments involved, and with organisational changes after the formation of the new organisation. A third pilot, *Natural Language Map* was abandoned because of difficulties securing access to the city's Geographical Information System.

2.8.5. Nisko Consortium

Nisko's role in the pilot evaluation was primarily to learn from the pilots in other cities and to support those where feasible. The Nisko Consortium represents a consortium of Polish towns and cities, mainly Nisko itself and Stalowa Wola. At the beginning of the project it was felt that the technical and legal infrastructure was not at a sufficient level to allow an EDEN pilot, and nor was there a sufficient level of Internet access.

2.9. Summary of the Assessment

2.9.1. Retrieval Performance: Finding Relevant Answers

The tests carried out in Bologna, Bremen and Antwerp are described in more detail in D6.2and summarised here. The test results for *Answer Tree* were produced with the assistance of Nisko, who provided a script to extract queries and results from log files, according to Napier's specification.

Address Guesser

If it performs satisfactorily the Address Guesser should respond to a user's query with a small number of office email addresses (and preferably only one) to which the user's message can be sent with confidence that the office is capable of answering the query. For validation and evaluation purposes this was expressed as the following targets:-

Validation: For test queries compiled by the pilot site project team, the relevant office should always be among those 'guessed' by the system, i.e. *mean recall should be 1.0*. Also a correctly guessed address should appear first among those guessed, more often than not. Otherwise the second one should be correct. This corresponds to an *MRR of 0.75*.

These targets were set very high because we had no comparable figures from the existing system. The current *Customer Satisfaction Service* mail routing system works differently; it does not give the user feedback on the destination office for a query, so there is no basis to calculate the proportion of queries that are routed correctly except by asking the participating offices to evaluate messages they received over a period. This task that was not feasible and, in any case, test results would not be comparable since the CSS tool is deployed for different offices.

Evaluation: For queries originating from citizens themselves, the targets were lower given that the NLP parser could not be expected to perform well with queries unless these were expressed in correct grammar. The revised targets were that a correctly guessed office should almost always be one of the first two shown, and at least half of the correct addresses should be shown. This corresponds to an MRR of 0.5 for 90% of queries (or 0.45), and 50% mean recall or mean average precision.

The results were as follows:-

	Mean Reciprocal Rank	Mean Recall	Mean Average Precision
Validation Targets	**0.75**	**1.00**	N/a
Run 1 (March 03) *	0.38	0.48	N/a
Run 2 (Dec 03)	0.40	0.49	0.38
Evaluation Targets	**0.45** or 0.50 for 90% of queries	**0.50**	**0.50**
Dec 03	0.35	0.37	0.35

Table 2.2: Address Guesser Performance

* Results previously included in D6.1 (Interim Tool Validation)

As a first remark we should note that we can only make reliable quantitative comparisons between runs if the change in factors between them is *either* some aspect of the system *or* some aspect of the test collection, but not changes in several aspects at once. Between the validation runs there were changes to the test queries and expected results as well as updates to the Answer Tree software and grammar. This means that we may only comment on the gap between performance and target, rather than say whether run 2 showed a percentage difference in performance from run 1.

In both the validation runs and the evaluation run (using citizens' queries) the performance was disappointing, especially since measures were taken after the first run to further 'tune' the system for better performance. The explanation is likely to be a combination of more than one of the following:-

a) The office descriptions were not detailed enough to be helpful in discriminating one from another, in relation to the queries.

b) The sample emails were not representative enough of the messages that each office should be able to respond to according to URP.

c) The expected results that were specified were not representative of the sample messages selected for training by the offices concerned. (i.e. the converse of (b))

d) The citizens' queries were not syntactically well-formed, so the NLP parser was unable to correctly match them on the basis of syntactic features shared with the office descriptions and training material.

Answer Tree

Answer Tree was validated in Antwerp, Bologna, and Bremen using the Dutch, Italian and German NLP grammars respectively. Then its performance in handling citizens' queries was evaluated in Bologna.
A user of *Answer Tree* may browse the FAQs it contains, which are hierarchically organised into 'groups' of questions, each of which has one corresponding answer that the user can read by clicking on the answer part. The ease of navigating this 'tree' was assessed through questionnaires (see section 2.9.2 below), but it is the search functionality that we are concerned with here.

The *Answer Tree* may be searched by the user by entering a question as a phrase, and this phrase is then used by the Answer Tree tool to retrieve 'related' questions. The key issues are whether all relevant answers are retrieved (i.e. effective recall), and whether the retrieved question/answer pairs are relevant to the query (i.e. mean average precision or mean reciprocal rank score).

Validation: For the first validation tests (late 2002) we considered that Mean Reciprocal Rank (MRR) was an appropriate summary measure because of its stress on weighting results that are displayed first. We also used mean recall as a secondary measure. We set the mean recall target at a very high 0.95

The reciprocal rank score was assessed for each of the 100 queries, then the mean calculated. Each query gives a set of question-answer pairs in response, and the first 4 pairs were assessed to judge whether each was an acceptable answer to the query. If any question-answer was acceptable, it's rank order in the results display was noted (e.g. 1st, 2nd ..). Then the reciprocal rank score (Radev et al, 2002) is the sum of the reciprocal ranks of all the acceptable responses. For example, if a query gives a set of 3 question-answer pairs, and the second and third are acceptable, the score is as follows:-

Q-A 1 0
Q-A 2 1/ 2 = 0.5
Q-A 3 1/ 3 = 0.33
Reciprocal rank = 0.83

The total possible score is 1/1 + 1/ 2 + 1/3 + 1/ 4 = 2.08. If there are no relevant answers, the score is 0 for that query. The target proposed in the Annex to D2.2 Addendum was for 75% of responses to include one acceptable result as the first question-answer shown, and for the rest to have one acceptable answer shown second, i.e. *a mean score of 0.87*.

In the second tests (late 2003) we added Mean Average Precision since (as remarked earlier) this is a more widely accepted measure, it includes an element of recall within it (thus avoiding having to interpret separate precision and recall figures), and also weights results that are shown at the 'top' of the results list more than those shown later.

For the validation tests 100 queries were compiled from the FAQs (or 'question-answer pairs'), and a note made of all that were relevant to each query. An important point to remember here is that, as in most question-answering systems, the number of relevant results that was expected was in most cases only 1 or 2, since FAQs are intended to be structured so as to provide the document that categorically answers the target audience's question (rather than a list of more or less relevant results).

When compiling the test queries, a balance was sought between: -

a) Queries worded to exactly match a phrase in the question or answer part of the FAQ
b) Queries worded to use the same words but in a different order and/or different forms of the same words (e.g. verb tenses).
c) Queries worded to use synonyms of words in the question or answer. This was of particular interest to Antwerp and Bremen.

Another key point to remember here is that *Answer Tree's* 'natural language' processing capabilities are built on top of a text retrieval algorithm, the open source SWISH indexing and retrieval mechanism, that is commonly deployed precisely because it is effective at retrieving exactly matching words and phrases. The NLP parser works invisibly to the *Answer Tree* user, in effect doing what an experienced human searcher would do to transform their search topic into a set of keywords, combining them where necessary with Boolean operators. So it was important to be able to check the effect of 'switching the parser off' to compare the results that could be obtained using a typical conventional search mechanism. This was particularly important given the lack of any existing FAQ search systems that could be used as a performance baseline.

Results

Before we discuss the results of the validation and evaluation tests, we should point out the accepted practice when interpreting differences between test runs. Referring to the Text Retrieval Conference (TREC), Buckley and Voorhees (2000) say that: -

> "One of the functions of the TREC conferences is to be a venue for discussions of what constitutes good IR experimental methodology. Simplifying enormously, the general consensus within TREC has been that Average Precision is a suitable evaluation measure for general purpose retrieval; that 25 topics is just barely enough for an experiment but that 50 topics is stable; and that 5% differences are worth noting" (pp.39).

We should point out that in TREC the test runs are judged on a sample of 100 documents taken from a much larger pool (Voorhees and Harman, 1999), whereas in the EDEN tests the collections are on fewer FAQ documents; 58 in Bologna and 53 in Bremen.

	Mean Reciprocal Rank	Mean Recall	Mean Average Precision
Validation Targets	**0.87**	**0.95**	N/a
Antwerp			
Run 1 (Dec 02) *	0.33	0.48	N/a
Run 2 (Dec 02) *	0.85	0.87	N/a
Bologna			
Run 1 (Dec 03)	1.21	0.86	0.81
Bremen			
Run 1 (Dec 03)	0.62	0.51	0.49
Run 2 (Dec 03) 'Parser on'	0.72	0.56	0.55
Run 3 (Jan 04) 'Parser off'	0.58	0.49	0.41
Run 4 (Jan 04) 'Syn. Handling'	0.67	0.54	0.50
Difference with tuning (run 2/1)	+17%	+10%	+11%
Difference with NLP (run 2/3)	+25%	+15%	+33%
Difference NLP+ syn. (run 4/2)	-7%	-4%	-8%
Evaluation Targets	**0.45** or 0.50 for 90% of queries	**0.50**	**0.50**
Bologna			
Run 1 (Dec 03) 'Parser on'	0.65	0.43	0.35
Run 2 (Jan 04) 'Parser off'	0.41	0.42	0.21
Difference 1/2	+24%	+2%	+14%
Bremen			
Run 1 (Jan 04) 19 queries only	0.58	0.47	0.43
Run 2 (Jan 04) 'Parser off'	0.26	0.24	0.16
Difference with NLP (run 1/2)	+123%	+97%	+171%

Table 2.3 Answer Tree Retrieval Performance

The results show four main features:-

1) The Bologna *Answer Tree* performed better than expected in responding to citizens' queries, with an approx. 2/3 chance that a relevant FAQ was shown first, although it did not reach the target for mean average precision.

2) The NLP parser was shown to make a material difference to performance, for the German and Italian parsers, as compared with the underlying SWISH text retrieval mechanism. This applied to queries compiled for validation testing and to queries entered by citizens themselves.

3) The synonym handling function when tested with the Bremen *Answer Tree* did not increase performance.

4) Performance in retrieving results relevant to the validation queries was quite dramatically improved by measures taken to tune the system, especially by refining the 'stop word' lists.

Citizens' Queries: The performance of the NLP parser exceeded the target for MRR, although it was below the target of 50% Mean Average Precision. It was also much higher than retrieval based on SWISH (using the same stop word list in both cases). The Bologna test was based on 49 queries from a sample of 100 taken from the log files, where the 49 represented queries where the Bologna PA officers judged that at least

one relevant answer existed. The Bremen test however may exaggerate the difference since their *Answer Tree* had not been piloted for long enough to gather a sufficient number of queries for reliable testing.

The NLP Parser vs SWISH. The comparison using the German grammar showed a convincing improvement in performance with the NLP parser than without. This test was based on a set of 100 queries that was evenly balanced between those that contained an exactly matching string (matching either the question or answer part of the FAQ), and those that varied the wording.

Synonym Handling: The mean average precision was noticeably *less* with synonym handling, demonstrating that this capability was better at increasing the proportion of non-relevant results retrieved than it was in retrieving additional relevant ones. The results need to be treated with some caution though, because the number of queries that actually included synonyms (20) was not enough to be confident of the effect.

2.9.2. From Access to Comprehension and Navigation

To summarise the evaluation of the pilot results on the criteria of improved access, comprehension and navigation we have used three categories:-

Fully Met:	The target was met in each pilot site where it could be assessed.
Partially Met:	The target was not met in one or more pilot site, but the results for the indicator were nevertheless positively evaluated by the site(s) concerned.
Not Met:	The target was not met in any pilot site.

The methods used to gather data on which to base the assessment of the indicators have been outlined in section 2.6. Further details are given in this section and in Chapter 3.

The indicators used are shown in tables along with the sources used and the outcome of our assessment. Each table is followed by discussion of the assessment. In this section the assessment is shown for each tool. In the next section we summarise the assessment of the tools acceptance, which is shown ordered by pilot site.

Address Guesser

Address Guesser was piloted in Bologna with the participation of the 10 municipality departments (or 'sectors') whose remit includes urban planning matters. The pilot was from 17 November to 14 December 2004, during which citizens were invited to try out the tool (and Answer Tree) by visiting a page accessible from the *Iperbole* civic network home page. The invitation was by email to the (approx. 18.000) subscribers to *Iperbole*, and promoted the tools as an experiment.

The municipality's Call Centre also tried *Address Guesser* for the duration of the pilot. Although Call Centre operators were not the main target group (ordinary citizens unfamiliar with the appropriate offices to contact for their enquiry), the operators are intermediaries for this target group, since their main role is to provide that information to telephone callers when the operators are unable to answer the enquiry themselves. So the tool was of potential value to them, as a means of looking up an office that they could advise a caller to contact (by any and all available methods). The Call Centre operators gave their views by recording how useful the Address Guesser was on those occasions, and through questionnaires and interviews by the EDEN team.

A questionnaire was also completed by 'back office' staff of the Office for Citizen Relations (URP) who handle emails 'routed' to the office through Address Guesser, or forwarded to them by other sectors who received the message in error.

Criteria	Indicator/ Target	Sources	Outcome
Access	- More than 50% of users say they 'do not usually get involved' in communication with the PA about planning.	Question-naire.	**Fully Met** Bologna: 66% of responses (n=70).
Access	- A decreasing % of messages get routed to the 'default office'	Question-naire, Interviews	*Insufficient data*. Only 3 messages in this category during the pilot.
Navigation	- Users de-select correctly guessed offices less often than they de-select incorrectly guessed offices.	Log files	**Not met** 39% of users de-selected a relevant office when one or more was shown, compared with 14% who de-selected a non-relevant one.
Navigation/ Com-prehension	- Fewer than 25% of users report serious problems.	Question-naire; Interview	**Fully Met** 4% of respondents disagreed they could work out what to do next without help.
Com-prehension	- Officers can answer routed messages as easily as direct e-mails to office.	Question-naire; Interview	**Not met** Routed messages were thought to need clarification more often.

Table 2.4 Assessment of Indicators for Address Guesser

The questionnaire responses to *Address Guesser* from citizens were more positive than expected given the disappointingly low retrieval performance described in the previous section. The tool is unlike the currently installed 'automatic routing' software in that (among other respects) it gives the user the opportunity to confirm (or de-select) any of the list of offices 'guessed' on the basis of what the user has typed. The analysis of their comments on 'how the tool can be improved', and of log files for the tool, suggests that this feature needs improving. In effect it gives the user the impression that there is uncertainty about the software's "guess".

A more general risk identified early in the project was that citizen-users would be unable to formulate enquiries properly without first knowing who they were addressing. The effect would be that the officers' handling the messages received would be unable to answer without clarifying what exactly it was that the user wanted to know. This appears to be borne out by the responses of URP officers who thought the enquiries were more difficult to answer without further clarification. Again this might be addressed by providing users with more information on the offices 'guessed' by *Address Guesser*.

Answer Tree

Answer Tree was piloted in Bologna from 17 November until 14 December 2003, using FAQs written for the purpose, and dealing with queries about: -

- Planning issues for business start-ups
- The Strategic Structural Plan (43 FAQs)

These were aimed at both Bologna's main target groups: respectively 'citizens in general' and planning professionals or others with an interest in the detail of city planning.

The *Answer Tree* tool was also piloted by the Bologna Call Centre operators, who currently refer to an online database of 'information sheets' (using Boolean searches), to help them answer callers' enquiries. Their views were recorded on whether the Answer Tree was effective in that role, for calls related to urban planning. The Answer Tree 'Administrator', an officer of the URP (Office for Citizen Relations) responsible for dealing with emails from users unable to find a relevant FAQ, also gave her views on the nature of these emails.

The Antwerp and Bremen partners were also scheduled to pilot *Answer Tree*, and both installed the tool and validated it. Antwerp (Digipolis) had a negative reaction from the Antwerp council department responsible (the Leien Information Point, responsible for handling citizen enquiries about a major re-development of the Leien, a road and thoroughfare through the centre of Antwerp). The role of Answer Tree 'administrator', to respond to messages from users unable to find a relevant FAQ, entailed more effort than was anticipated and the pilot was withdrawn before further evaluation data could be gathered.

Both Antwerp and Bremen were able to carry out usability tests after first validating the tool, but with only 3 and 2 participants respectively. Both sites had difficulties recruiting participants, and in Bremen's case the fine-tuning and validation of the tool took longer than anticipated and there was insufficient time to gather further data. More details of their validation and usability tests are given in Chapter 5.

Criteria	Indicator/ Target	Sources	Outcome
Access	1. More than 50% of users say they 'do not usually get involved' in communication with the PA about planning.	Question-naire.	**Fully Met** Bologna: 75% of responses (n=24).
Navigation	2. More than 50% of users satisfied with FAQ shown.	Log files, from satisfaction question built into tool.	**Not Met** 6% of users were fully satisfied and 4% satisfied, but 15% were not satisfied and 72% did not give any satisfaction rating
Navigation	3. Fewer than 25% of questions forwarded to Admin already in Answer Tree	Question-naire; Interview	**Fully Met** None were in this category, although only 4 were received.
Navigation/ Com-prehension	4. Fewer than 25% of users report serious problems.	Question-naire; Interview	**Fully Met** 4% of respondents disagreed they could work out what to do next without help.

Table 2.5 Assessment of Indicators for Answer Tree

The citizen questionnaire and Call Centre operators' views were again the main source of the Bologna evaluation, and the responses were considered promising. The encouraging results on the retrieval tests described in the previous section of this chapter appeared to be matched by favourable views from citizen users, who reported few serious difficulties (i.e. those that stopped the user progressing without help) with navigation or the clarity of the results. This was also the case with Bremen's usability field testers, although there was criticism of the 'satisfaction rating' feature. In Antwerp there was one usability issue that was considered serious: test users with little Internet experience did not have sufficient screen cues to indicate they should click on a question to see the answer.

The Bologna questionnaire response rate was disappointing, as was the response to 'satisfaction ratings' that users had the opportunity to give after seeing the results page for their query. The ratings themselves (10% satisfied, 15% not) were promising given that this was a prototype service, but the 75% non response

rate suggests a design flaw (as do the Bremen usability tests). The indications are that it was wrong to associate the 'satisfaction rating' with the results page, and instead it should have been associated with each answer since users were unwilling to back track from the answer to the results page in order to give a rating.

Guided Forum & Notification Handler

The tool was implemented twice in Bremen. The outcomes of the first pilot in the Horn Lehe neighbourhood of the city were described in deliverable D6.1, and the second in Chapter 5 of D6.2. The Antwerp pilots were unsuccessful for reasons given earlier. The *Notification Handler* was not used in the pilots, but for consistency the targets that were agreed for the tool are shown below.

Criteria	Indicator/ Target	Sources	Outcome
Access	1. More than 50% of users say they 'do not usually get involved' in communication with the PA about planning.	Question-naire.	**Partially Met** Most users agreed but there were only 19 responses.
Access	2. A higher % of the consultation responses are received via the forum than other channels.	Records of consultation responses	**Fully Met** Almost no contributions by non-EDEN channels.
Access	3. The forum attracts more participants as % of neighbourhood population, than previous pilot.	Forum responses	**Not met** Response was lower than previous pilot.
Access	4. Increasing % of page requests to forum are from notification emails	Web server logs	**No data** Functionality not deployed.
Navigation	5. Increasing % of forum visits include requests to information pages	Web server logs	**Partially Met**
Navigation/ Com-prehension	6. Fewer than 25% of forum users report serious problems.	Question-naire; Observation.	**Fully Met**
Navigation/ Acceptance	7. Higher % of forum users post more than one message, than in previous pilot.	Forum responses	**Not Met** Few repeat postings

Table 2.6 Assessment of Indicators for Guided Forum & Notification Handler

Since the 'owners' of the e-consultation considered that its circumstances and high level of responses relative to other channels compensated for the low response relative to other fora, the value of the pilot is not fully reflected in the quantitative indicators. This value is discussed further under "Acceptance" below.

Natural Language Map

The tool was implemented in Antwerp and Bremen, but unfortunately there was insufficient time to carry out any validation or evaluation of it. For consistency the targets that were agreed for the tool are shown below.

Criteria	Indicator/ Target	Sources	Outcome
Access	1. Page requests to related documents increase relative to search page ('browse-to-act ratio')	Web server logs	No data
Navigation	2. More than 50% of users satisfied can find a relevant answer online	Web server logs	No data
Navigation/ Com-prehension	3. Fewer than 25% of forum users report serious problems.	Questionnaire; Observation.	No data

Table 2.7 Assessment of Indicators for Natural Language Map

Style Enhancer & Multi-Language Helper

The *Style Enhancer* evaluation involved planning professionals associated with the Archivio Osvaldo Piacentini, who also recruited citizens to assess the effects of 'style enhancing' on their comprehension of documents on urban planning. The *Multi-Language Helper* module could not be deployed in time to carry out any evaluation of it, partly because of issues in identifying who should test it. This is discussed under 'Acceptance' below.

Criteria	Indicator/ Target	Sources	Outcome
Navigation/ Com-prehension	1. Fewer than 25% of users report serious problems.	Question-naire.	**Fully Met** No serious problems reported.
Com-prehension	2. Professional users give higher 'confidence ratings' on readability of documents after/before checking & translation	Forum responses	**Fully Met**
Com-prehension	3. Citizen readers give higher ratings of readability of documents after/before checking & translation	Question-naire;	**Partially Met**

Table 2.8 Assessment of Indicators for Style Enhancer & Multi-Language Helper

The target users of the *Style Enhancer & Multi-Language Helper* tools are mainly planning professionals working on documents to improve their 'readability' or comprehension by ordinary citizens. Citizens should benefit, but indirectly rather than as users themselves. Note that the tools should be usable in combination, i.e. after uploading a document and checking the phrases highlighted by the *Style Enhancer*, if the Multi-Language Helper tool is available the user can choose to include translations of selected phrases (and target languages) from the glossary used by both tools.

For the evaluation, testers were recruited through the *Archivio's* subscribers and contacts, 6 of whom were invited to try out the tool on documents of their choosing. They were asked to rate their confidence that the documents were more readable after using the tool to 'enhance' it. Then 12 citizens were asked to give 'readability ratings' to each of 3 documents (selected by one of the professional users).

From the users' (professionals) evaluation:-

- No serious usability problems were encountered, the language used was thought easy to understand, and the system response time quick enough.
- After checking the documents with SE, professional users' confidence in their documents' readability for the target audience increased, mostly to 1 point above the median on a 5 point scale. In their questionnaire responses, half the testers agreed that, after using SE, they could be more confident their document would be easily understood. The other participants were 'neutral' on this point.
- The main feature that some testers felt needed improvement was the scope of the glossary. One tester thought the syntax checks should be more rigorous, and another that the SE site should include links to further sources of help for authors of urban planning documents.

From the citizens' evaluation:-

- The participants' readability ratings for the 'style enhanced' documents increased by 1 (39% of ratings) or 2 (6% of ratings) points on a 1-5 scale, however they were mostly unchanged from the un-enhanced documents (56% of the ratings).
- Where the 'style enhancements' were rated more readable, this was despite large differences between the participants in terms of their readability ratings, the particular expressions each person found difficult, and the number of those expressions and the reasons each was 'difficult'.

2.9.3. **The Tools' Acceptance and Impact on Participation**

In this section we outline the indicators for acceptance of the tools and how these were assessed on the basis of each pilot. Then for each pilot we discuss the impact on participation, in terms of changes in the various practices that the tools were intended to support.

Bologna: Address Guesser and Answer Tree

The overall assessment of the Answer Tree and Address Guesser pilots by CoBo took account of four main factors: -

1. The impact on enquiry handling seen in the pilot and the likely impact in the future,
2. The questionnaire responses from citizens who tried the tools,
3. The usefulness (or potential usefulness) of the tools for the Call Centre function
4. The effort needed to maintain each tool

The Impact on Enquiry-handling: Although the service was 'experimental' the volume of enquiries received by URP was comparable with that by each of the more conventional channels in the pilot period. In terms of the volume of enquiries during the pilot that impact was small (around 8 enquiries per channel). For *Answer Tree*, CoBo would expect benefits in terms of a decrease in queues at the Front Office/URP desks, and in Call Center calls regarding the specific urban planning topics (i.e. General Master Plan) that the FAQs concerned. Thus, officers could invest more time on other tasks. CoBo believed the % of benefits for Front Office/URP desks and Call Center will depend on the broadening of the domain from Urban Planning. If that task is undertaken, the % of benefits could be around 15% of queues at Front Offices.

Usefulness for Call Centre Operators: The operators involved gave a favourable view of the potential of both tools, although they were generally not useful for Call Centre operations during the pilot period. The operators thought the *Answer Tree* FAQs to be less focused on the kinds of question they answer by telephone than were the Information Sheets currently retrieved from an online database. The main drawback of *Address Guesser* was that it did not provide contact details (other than email) for the offices concerned.

Questionnaire responses: These were favourable on almost all the questions asked about the usability of the tools and overall satisfaction. The profile of the respondents was encouraging, in that users were mainly people who do not normally contribute their views to the municipality on planning matters..

	Indicator/ Target	Sources	Outcome
Address Guesser	1. More than 50% of target users say they are satisfied with their visit.	Question-naire.	**Fully Met** 73% of target group satisfied, and same for all respondents (*n=70*)
	2. Call Centre Operators find the tool helpful in identifying offices for forwarding enquiries to.	Call Centre logs; Interviews.	**Partially met** Not met on basis of call logs, but operators attributed this to content (lack of contact info).
	3. 95% of visits result in a message being sent	Log files	**Not Met** 60% guesses led to confirmed message.
	4. More than 50% of users say the site better prepares them to comment on what the council is doing in their neighbourhood.	Question-naire.	**Not Met** 47% agreed.
	5. Less effort for PA officer to add new office, compared with CSS	Discussion	**Partially met** Fine-tuning needs attention to content and distribution of message samples used for training the system.
	6. Pilot site rating of maintenance effort/ added value.	Discussion	**Partially met**
Answer Tree	1. More than 50% of target users say they are satisfied with their visit.	Question-naire.	**Fully Met** 52% of target group satisfied, 60% of all respondents (*n=24*).
	2. Call Centre Operators find the tool helpful in answering questions.	Call Centre logs; Interviews	**Partially met** Not met on basis of call logs, but operators attributed this to FAQ writing style
	3. More than 50 % of direct questions to PA relevant for online publication	Interview AT admin	**Fully Met** 80% considered relevant
	4. More than 50% of users satisfied with ability to find answers quickly enough.	Question-naire.	**Not met** 34% agreed, 38% disagreed.
	5. More than 50% of users say the site better prepares them to comment on what the council is doing in their neighbourhood.	Question-naire.	**Not met** 38% agreed, 38% disagreed.
	6. Pilot site rates added value higher than maintenance effort	Discussion	**Partially Met**

Table 2.9 Acceptance of Address Guesser & Answer Tree in Bologna

Maintenance effort: *Answer Tree* maintenance was considered a relatively easy task, mainly comprising the online updating of the FAQ tree, and periodic indexing of the content. The 'satisfaction ratings' provided by users were disappointing, and additions of further 'plain Italian' content to give more exhaustive coverage is considered a prerequisite for increasing performance. *Address Guesser* maintenance was thought more problematic. Effort may be saved relative to the current 'automatic routing' module of the Customer Satisfaction Service, since keywords do not have to be defined. However the collection of training material for *Address Guesser* was the most resource-consuming task of the overall project.

Bologna: Style Enhancer & Multi-Language Helper

	Indicator/ Target	Sources	Outcome
Style Enhancer & Mult-language Helper	1. Users rate 90% of advice messages and translated phrases as useful and understandable.	Content samples, interviews	**Partially Met** 67% on average
	2. Pilot site rates added value higher than maintenance effort	Discussion	**Not Met**

Table 2.10 Acceptance of Style Enhancer & Multi-Language Helper in Bologna

The acceptance by *Style Enhancer* test users was quite positive, but not enough for the pilot site (CoBo) to justify further deployment. The main consideration here was that the documents that it proved more useful for were also those that the tool was least needed for, i.e. short documents already written for a general audience, rather than longer technical documents where the need for 'enhancement' was greater. The *Style Enhancer* was not thought to add anything to the URP's existing skills and capacity to edit short documents for a general audience. On the other hand it was felt to be promising for specialists/ planning professionals who write documents that should also be understandable for the general public, despite some indications that the tool was not at a sophisticated enough stage for that purpose.

Bremen: Guided Fora, Answer Tree and Natural Language Map

	Indicator/ Target	Sources	Outcome
Guided Fora	1. More than 50% of target users say they are satisfied with their visit.	Question-naire.	**Partially Met** 54% agreed, but there were only 14 responses.
	2. More than 50% of forum users feel more involved in decision-making.	Question-naire; Interview	**Partially Met** 39% agreed, but there were only 14 responses.
	3. Ratings of contribution quality higher than previous fora.	Content analysis; Interview	**Partially Met** Most contributions considered good, although overall ratings lower than previous fora.
	4. Pilot site rates added value higher than maintenance effort	Discussion	**Partially Met**
	5. More than 50% of Notification Handler users are confident in privacy of user profiles.	Question-naire;	No data Function was not deployed.

Answer Tree	6. More than 50% of target users say they are satisfied with their visit.	Question-naire.	Insufficient data Pilot not long enough
	7. More than 50 % of direct questions to PA relevant for online publication	Interview AT admin	Insufficient data Pilot not long enough
	8. More than 50% of target users satisfied with ability to find answers quickly enough.	Question-naire.	Insufficient data Pilot not long enough
	9. More than 50% of target users say the site better prepares them to comment on what the council is doing in their neighbourhood.	Question-naire.	Insufficient data Pilot not long enough
	10. Pilot site rates added value higher than maintenance effort	Discussion	**Partially Met** On basis of user requirements, validation, and expectations.
Natural Language Map	11. More than 50% of target users say they are satisfied with their visit.	Question-naire.	No data
	12. More than 50% of target users say they are more confident they know what plans affect area(s) they are interested in	Question-naire.	No data
	13. Pilot site rates added value higher than maintenance effort	Discussion	**Not applicable**

Table 2.10 Acceptance of Answer Tree, Guided Forum & Natural Language Map in Bologna

The Waller Heerstrasse Guided Forum pilot was the second in Bremen, and in a quite different environment from the initial pilot in the Horn Lehe neighbourhood. The two neighbourhoods differed in socio-economic status, Waller Heerstrasse pilot being economically poorer. The consultations also differed in that the first (Horn Lehe) was an *informal* debate on the future of the area, while the second (Waller Heerstrasse) formed part of the *formal* early participation phase of a zoning code procedure. This entails collecting citizens' views as a supplement to other legally required means of participation (like a public meeting and the ability of citizens to join a meeting of the district committee where the issue is handled and the opinion of the district committee becomes an official part of the procedure.

The Waller Heerstrasse pilot was considered successful in terms of the quality of the contributions, despite getting relatively few of these and an even lower response to the online evaluation questionnaire. The response to the forum was higher than by conventional channels, and expectations were not high since the issues for discussion were not controversial.

Critical issues affecting the forum pilot were:-

- Motivation of PA staff involved: It is absolutely necessary that PA-members working close on the problems discussed in a forum participate actively by responding directly to citizens' requests. One-way-questions and contributions with no direct responses make the user feel they are not taken seriously.
- A political commitment of the institutions in charge of the issue discussed is needed. This does not mean that all comments are followed up, but it must be communicated that all comments are taken into account.
- The resources involved in setting up the pilot.

For further deployment in Bremen, according to central PA officials in charge of master planning, the use of guided fora can make sense in projects with citywide relevance. Deployment of the *Guided Forum* tool is not only suitable for formal building and planning procedures but also within other (informal) procedures like discussion of a "regional concept" or the pavement of the central market place. Apart from fora in restricted

time frames (like four to five weeks in the two EDEN-pilots), also longer less time-specific fora seem to make sense. Further deployment of the *Answer Tree* tool was thought likely, even although this tool was not originally selected for piloting in Bremen. The tool was expected to meet a long-term requirement for accessible background information on the Waller Heerstrasse web site, and the validation results were positive. Unfortunately the installation and tuning of the tool took longer than anticipated and it could not be piloted for long enough to get any useful data from citizens' usage of it.

2.10. Conclusions and Recommendations

2.10.1. Deploying NLP for Public Participation: Results and Further Work

The pilot sites each had a desire to improve their online capabilities in the area of citizen engagement, but with preferences for different EDEN tools, and quite different deployment contexts. There is nothing in Natural Language Processing technology that makes it an inherently suitable tool for citizen engagement, and there are other measures that would help to accomplish that aim. The key assumption or working hypothesis of EDEN however has been that NLP can 'make a difference' when deployed for well-defined purposes as part of an infrastructure (human and technical) meant to support citizen engagement. There are two aspects to the underlying logic of EDEN; firstly that NLP may reduce the effort needed by citizens to find relevant answers to their questions and understand them when they find them, and secondly that NLP may reduce the administrations' effort in handling the more routine communications involved in providing a response to citizens' concerns.

The pilots met their objective of demonstrating that the NLP approach deployed in EDEN can reduce the barriers citizens and PAs face in communicating online, and in doing so may encourage those citizens who do not normally take part in city planning to contribute their views through online channels. In particular: -

1. The NLP approach implemented in *Answer Tree* was shown to be better at retrieving relevant FAQs in response to natural language queries than the widely used SWISH indexing and retrieval algorithm.
2. The *Address Guesser* tool showed promising results in finding relevant PA office addresses by comparing users' queries with those previously answered by them. Although the results were not accurate enough for users to be confident that the 'guessed' addresses were correct, refinements to the 'training' samples used and to the interface design appear likely to meet that objective.
3. The *Style Enhancer* tool was considered useful by planning professionals, particularly for checking relatively short documents giving general information to citizens. The tool is likely to be effective as a complement to a human editorial function although it is unlikely to replace that role. Further development of the glossary to differentiate between domain –specific and general usages of words and phrases would enhance the tool's effectiveness in that role.
4. The users who tested the tools, on a self-selected basis, were satisfied with them and were mostly people who normally make enquiries by telephone or in-person. This indicated a potential uptake of online enquiry-handling estimated at 15% with wider deployment across other PA sectors than urban planning, although the short length of the pilots did not allow sufficient volumes of enquiry data to be used in making this estimate.
5. The pilot users (citizens) were mostly people who do not take part in city planning consultations by the traditional means (e.g. public meetings). Sizeable minorities of them agreed that the tools better prepared them to contribute their views online.
6. The *Guided Forum* pilots demonstrated that the recruitment of local citizens to help moderate online discussions can also help publicise neighbourhood or district level consultations to the citizens affected by planning decisions. The Bremen pilots also demonstrated acceptable levels of contribution quality and higher response rates than the traditional means, despite their small scale.

On each of the main criteria (access, navigation, comprehension and acceptance) some results were below expectations. There were also some objectives that were unmet, and which we believe deserve further work to address. We summarise each of these aspects of the pilots below.

To get satisfactory results from the NLP 'enquiry handling' tools (Address Guesser and Answer Tree) citizens need to express their query without grammatical or spelling errors, and formulate the topic in enough detail for a good match to be made with the relevant information. None of these conditions are realistic for a fully deployed system, and further work would be needed to implement strategies to address them. For example:-

- The addition of a spell-checking feature.
- The use of a synonym-handling to broaden or narrow searches needs to be examined in more depth. Although implemented and tested in EDEN the results were inconclusive.
- The relaxation of the syntactic rules used to parse queries and the addition of rules to handle specific kinds of 'badly formed' syntax could provide better results with everyday language, but would need detailed work to establish how far such rules should be relaxed without compromising performance on queries that *are* 'well formed'.

The NLP approach is based on the analysis of syntax, and as such does not claim to be fully descriptive of language since it does not address its semantic and pragmatic aspects, i.e. (respectively) what different words mean in relation to each other (as represented in a thesaurus for example), and how meaning is given by the context that language is used in (e.g. to identify what 'there' refers to in "I live in Bologna. Have you ever been there?"). At least the semantic aspects are addressable, through applied research on 'ontologies' that represent and manage such relationships. This could be used (for example) to improve the effectiveness of the *Style Enhancer* glossary, or to refine performance of *Address Guesser*.

More research is needed on how different 'genres' or forms of information provision serve the purpose of enabling e-participation. The standard 'relevance' measures of precision and recall are widely known to be insufficient for evaluating the utility of information retrieval tools, i.e. the practical relevance of the texts that match the users' queries. By 'practical relevance' we mean the utility of the information as a resource for accomplishing the users' aims (getting permission to erect a balcony for example, or lobbying a district planning committee about traffic conditions). In the evaluation of EDEN the (relative) novelty of the retrieval approach meant that a focus on the standard measures was necessary. Practical relevance was assessed, through satisfaction ratings and questionnaire responses, but in-depth enquiry into the uses made of the information was not feasible within the scope of the project. It would be worthwhile for example to compare the practical relevance of information provided in FAQ form (as in Answer Tree) with similar information provided in relation to online maps (as in Natural Language Map).

More research is needed into the relationship between citizens views *about* PA information provision and their attitudes towards e-participation. The assumption that more accessible professional-quality information about city planning is sufficient to encourage citizens to play a more active part was taken as given in the project, rather than treated as a topic of enquiry in itself. It would be particularly interesting to carry out a comparative study of such attitudes in political cultures that do and do not have a tradition of local consultation by PAs (e.g. Germany and Poland respectively).

The flexibility or adaptability of e-participation tools needs further investigation, to encompass the development methodology as well as the functionality of the tools. The EDEN Guided Forum tool, as specified in the project, had a range of features that there was no opportunity to test because although they were foreseen, they turned out not to be necessary for the circumstances of the pilots. This included features meant to support successive phases of policy making, opinion polling, and notification about consultation events. Also some relatively minor interface features that were adaptable were not easy enough to adapt, in effect because the project schedule followed a 'waterfall' approach (requirements – specification – implementation – testing) that provided limited opportunities for the prototypes to evolve according to a flow of events that (being political) was not within the control of the project. So: -

- The 'unused features' of the EDEN fora need further investigation in the context of phased policy making.

- E-participation tools need to accommodate the changing circumstances that are inevitable in a political environment, and development methodologies that give more focus to rapid, evolutionary, prototyping may be better suited to that.

All of the tools developed in the project would have benefited from longer pilot periods and that of course applies most to the two tools that were developed but could not be piloted in EDEN, the Natural Language Map and the Multi-Language Helper. Both present research issues and opportunities. *Natural Language Map* has already been mentioned above. *Multi-Language Helper* did not fulfil the anticipated needs in EDEN because of the languages implemented, rather than the design principle which remains relevant.

2.10.2. Strengths and Weaknesses of the Methodology

The EDEN project had ambitious aims that were challenging to evaluate. The aims of 'informed participation' and 'improved communication' are general expressions of political will, as laudable as 'knowledge-intensive production' or 'improved business performance' are in the commercial world, but with a much shorter history of evaluation practice with which to assess the changes brought about by technology. A wide range of approaches have been deployed in attempts to assess the less tangible impacts of technology on commerce in terms that can be correlated with quantifiable performance. E-democracy research has few equivalents, although there have been recent developments in the direction of cost-benefit analysis for evaluation of e-government services more generally (e.g. the Value of Investment approach[1]).

We regard it as strength of the approach that it has been flexible enough to address the changing circumstances of the project in such a way that its main elements (requirements, technology, plus evaluation criteria and methods) have been well defined, without compromising the research aspect of the project. In terms of the latter, we had to balance the need for rigorous application of research methods with the need for relevance. What follows below therefore starts with some comments on the balance drawn between rigour and relevance, and then considers the strengths and weaknesses of the evaluation methods.

Relevance, Rigour and Sampling Issues

The research methods have been applied by the EDEN teams in each of the participating Public Administrations by people whose background is generally *not* academic research (excepting Bremen), but providing technology-based services to citizens on behalf of administrations. As coordinators of the evaluation, we have needed to balance our own concerns and academic interests with the need for the PA partners to 'own' the process and outcome, in the sense of understanding and feeling committed to both.

On the other side of the 'balance', the evaluation would risk misleading the Public Administrations if it led them to conclusions that were not based on reliable evidence. We are confident that in EDEN the PA partners' piloting and deployment decisions have not been influenced by any evidence that they were not convinced about, or which over-rode their own judgement.

The corollary of that is our need as academic partners to demonstrate that proper care has been taken to gather convincing evidence, according to established research principles. The first two elements of the research approach discussed earlier, action research and ethnography, are established qualitative research approaches. Although they do not concern us here, the third element 'evaluation' is a label for a wide range of methods that are also based on widely variant assumptions. It is important therefore to have a focused set of principles to ensure validity. The standards and definitions of validity appropriate for case studies have been extensively discussed by Yin (1989), for whom validity encompasses:-

[1] IDA Value Of Investment (VOI) Final report v 2.1 available at: http://europa.eu.int/ISPO/ida/jsps/index.jsp

Construct validity
This is the establishing of correct operational measures for the concepts being studied. In EDEN these took the form of the evaluation criteria and indicators, and their validity was addressed using the three 'tactics' recommended for case studies:-

- Using multiple sources of evidence, or 'triangulation'. This is the rationale for adopting a wide range of methods to provide alternative sources of data for the assessment. In practice not all sources could be used within the available resources, which was why a wide range of indicators were proposed at the beginning of the work package and then narrowed down according to what was practicable.
- Establishing a chain of evidence: The principle here is to "allow an external observer…to follow the derivation of any evidence from initial research questions to ultimate case study conclusions" (Yin, ibid. p102). We believe that principle has been met in this report and interim reports that preceded it, underpinned by the electronic records of log analysis, questionnaires and other sources used.
- Participant review: Inviting participants to review research findings helps maintain construct validity since it reduces the likelihood of falsely reporting an event or misrepresenting people who have contributed their views. In EDEN that has taken the form of continued opportunities for city partners to comment on draft reports. However it has been a weakness of the project that it did not (and could not) establish and maintain closed user groups in each city, to consult with throughout the project.

*Internal validity***:** this requires "explanation building", an iterative process of making initial statements about the research data, revising the statement in light of new evidence, and continually seeking other "plausible or rival explanations" (Yin, ibid. pp. 113-115). In EDEN the focus has been on explanation of the acceptance (or not) of the tools by citizens and PA users and the impact on participation. The explanation building has been limited by the time available for analysis, and by the very practical focus of the evaluation. This is a weakness that should be remedied in future projects by longer pilot periods.

External validity is the basis on which generalizations are made. Yin notes that: -

> *"...The analogy to samples and universes is incorrect when dealing with case studies*. This is because survey research relies on statistical generalization, whereas case studies (as with experiments) rely on *analytic* generalization. In analytic generalization, the investigator is striving to generalize a particular set of results to some broader theory" (op.cit. pp.43-44, emphasis in original).

In EDEN the broader theory we aim to contribute to takes two forms; firstly, the 'best practice' literature that informs policy-making about online citizen engagement; and secondly the academic literature on ICTs and democracy. We should nevertheless emphasise that statistical generalisation is not possible from the questionnaire data presented in this report.

That leads us to comment on the *representative* nature of the views obtained from users. In usability research it is considered normal to engage 3 - 5 individuals for in-depth usability lab testing, while 6 – 9 people are typical for focus groups, and 30 is considered the minimum for prototype testing questionnaires (Nielsen, 1993). The accessibility of online information for e-participation purposes is not an area that is well researched, as we already noted. An approach based on statistically rigorous surveying methods, whether a controlled statistical experiment or simply a random survey sample, would therefore not have been effective or practical in our view.

It would have been desirable for many reasons including validity to recruit a closed user group at the outset of the evaluation, to allow in-depth questioning of the same group of people over time. We would strongly recommend that approach in future projects, although it was not possible in EDEN for a variety of reasons, primarily the resources needed to maintain communications and commitment, but also because of the difficulties in establishing early enough exactly which neighbourhoods would be involved.

Evaluating Retrieval Performance

The EDEN project took NLP technology 'out of the laboratory' and to assess its performance also meant applying information retrieval methods that are normally used in the laboratory. The measures used have

been standard ones, but applied in an unconventional way and with some relaxation of the controls normally used in laboratory settings.

Two important points about these measures are, firstly, that they are dependent on a more or less subjective judgement about the relevance of the results to the question that was asked, so the test results are dependent on the testers and the system being tested. The key element of control here is to vary only one element of the test collection or system in question, and it is a strength of the project that comparisons could be made on that basis given the constant pressure to improve all elements of the pilots.

Secondly, the tests are normally carried out by information retrieval specialists rather than by ICT staff in city councils, and normally using questions that are worded to match the capabilities of the system and the information being tested. This was a critical issue, since it highlighted differences in expectation between the testers (city partners) and the software suppliers (technical partners) about how and whether the NLP tools should process queries that were not grammatical enough to be 'natural language' in computational linguistic terms. As we pointed out earlier the targets for evaluation with real users' questions were lower, even though the internal validation gave excellent results. While it may seem perverse to treat the lowering of expectations as a *strength*, this was an indication of the greater awareness among the PA partners of what NLP was and was not capable of doing, which resulted from their active role in the validation of retrieval performance.

A *weakness* of these tests was that the targets set at the beginning of the validation were not based on a comparative analysis with current systems, for the good reason that there were no systems that were sufficiently similar to perform quantitative comparisons. This meant that the targets were rather arbitrary (and optimistic). The comparative analysis of the NLP parser in *Answer Tree* with SWISH was therefore a *strength* of the evaluation, although the resources to undertake it were stretched.

Web Server Log Data and Usability Testing

The use of these methods in the evaluation suffered from a strain on resources at the end of the project and sheer lack of time. In Bologna's case log files were available but could not be analysed in time to include the results in this report. Usability tests were carried out on a very limited scale in Antwerp and Bremen and were considered valuable by the project team members there, even although the numbers involved were not enough to be confident in the reliability of the results and the reporting was less rigorous than had been hoped for. We have no doubt that both methods would play an essential part in more extensive pilot tests.

Understanding Citizens' Experiences of E-democracy

The scope of the evaluation, its resources, and the multi-lingual nature of the consortium effectively ruled out the gathering of extensive *qualitative* data on how using EDEN tools changes citizens' experiences and views about the activities that EDEN is intended to support. That is, substantially more could be said about what constitutes 'making an enquiry', or 'understanding an urban planning document' or 'representing one's views about the city's plans', and how the experience of having more accessible information to hand changes the relationships between citizen, administration and elected representatives. To describe and explain that would require a different kind of project, involving longer periods of actual use, greater input of qualitative researchers fluent in the pilot city language(s), and the experience of implementation and deployment that the project has successfully gained.

2.10.3. Moving from e-Enabling to E-Participation

The evaluation was framed in terms of a 'trajectory from e-enabling to e-participation' since EDEN may be seen as an attempt to ground e-participation on the improved accessibility of information that could help citizens to take part in their local council's decision-making and/or ensure their own decisions comply with planning regulations. The evaluation sought to establish a connection between information accessibility and participation in decision-making, a task made more difficult by the choice of tools made by the pilot sites. The Guided Forum was not piloted in Bologna alongside the NLP-based tools, and the latter were not piloted in Bremen. Bremen did however invest considerable effort in providing 'professional level' information to

accompany each pilot of the Guided Forum and found the level of informed contribution to be higher than expected, even though evidence that contributors were informed *by the online information* was inconclusive.

The connection between information accessibility, i.e. the ease of getting relevant answers to questions (on the one hand) and making informed responses to public consultations (on the other hand) appears then to remain at an abstract level – neither proved or disproved as a result of the EDEN pilots.

However if we look at the experience of the pilots there *were* instances where citizens' active participation mediated by online access had material consequences for the pilots. Those instances were:-

- The citizen moderators who became engaged in the Bremen Guided Forum pilots amplified the capacity of the Public Administration to handle the additional 'channel' for consultation responses and was a pre-requisite for the success of the forum as a means for representing citizens' views. The success of that role depended however not so much on the professional level information provided by EDEN, but their access to online tools, existing familiarity with local people's everyday lives, and their ability to promote the forum to them.
- The response rate from citizens was noticeably and consistently higher in Bologna than in other pilot cities. This can probably be attributed to the consolidated base of subscribers to the Iperbole network, the Bologna Municipality's role in that network, and their ability to contact citizens directly (electronically). The response relied on the good will of Iperbole subscribers, their interest in its further developments and more specifically in the potential of EDEN, since no other incentive was offered for their participation. That participation, in the form of the online questionnaire responses and the queries entered in Answer Tree and Address Guesser, was a pre-requisite for working out how to improve performance for further deployment.

These examples fall short of the definition of 'e-empowerment' given earlier (the pilots did not demonstrate a 'bottom-up' influence of citizens' views on planning decisions). However they do indicate the dependence of e-enabling initiatives like EDEN on other online initiatives involving partnership with citizens, that are 'e-empowering' in the sense that citizens influence the design agenda and deployment policies that create the infrastructure for e-democracy. The trajectory of the EDEN pilots shows that the 'trajectory from e-enabling to e-participation' is not a linear one. Rather than inferring that online information-provision must be 'good enough' before further steps can be taken to e-empowerment, we should see the two as inter-dependent.

3. References

Buckley, C. and Voorhees. E. (2000) Evaluating evaluation measure stability. In *Proceedings of SIGIR '00*, pp. 33-40. ACM Press, July 2000.

Caroll, J.M and Rosson, M.B (1992) Getting round the task-artefact cycle: how to make claims and design by scenario. *ACM Transactions on Information Systems*. 10, pp. 181-212

Checkland, P. and Scholes, J. (1990) *Soft Systems Methodology in Action*. Wiley, Chichester, UK.

Fountain, J. (2002) Toward a Theory of Federal Bureaucracy for the Twenty-First Century. In: *Governance.com: Democracy in the Information Age*, Ed. E. Karmarck and J. Nye, Brookings Institute Press, Washington, D.C. pp. 117-140.

Macintosh, A. (2004) Characterizing E-Participation in Policy-Making *Proceedings of the Hawaii International Conference on Systems Sciences HICSS-37* January 5-8, 2004., Hawaii.

Macintosh, A., Davenport, E., Malina, A.; and Whyte A.(2002) Technology to Support Participatory Democracy. In *Electronic Government: Design, Applications, and Management*. Ed. Åke Grönlund. Idea Group Publishing, January 2002; pp. 223-245.

Nielsen, J. (1993) Usability Engineering, Academic Press, London.

OECD (2001). *Citizens as Partners: Information, consultation and public participation in policy-making*: OECD, Paris.

OECD (2003). *The e-Government Imperative*. OECD, Paris.

Radev, D., Libner, K., and Fan, W. (2002) Getting Answers to Natural Language Questions on the Web *Journal of the American Society for Information Science and Technology* 53(5), pp. 359-364

Spink, A. (2002) A user-centered approach to evaluating human interaction with Web search engines: an exploratory study. *Information Processing and Management* 38(3): pp. 401-426

Suchman, L. and Trigg, R. (1991) Understanding Practice: Video as a Medium for Reflection and Design In: *Design at Work: Cooperative Design of Computer Systems*. Ed. Greenbaum, J. and Kyng, M. Lawrence Erlbaum Associates, Hove, U.K.

Voorhees, E. and D. Harman, D. (1999) Overview of the seventh text retrieval conference (TREC-7). In *Proceedings of the Seventh Text REtrieval Conference*, 1999. NIST Special Publication. 197

Westholm, H. (2003) “Adaptability” in online democractic engagement: a multi-channel strategy to enhance deliberative policies. *Communications*. 28. pp.205-227.

Whyte, A. and Macintosh, A.; (2002) Analysis and Evaluation of e-consultations; *e-Service Journal*; Volume 2, No 1 “e-democracy in Practice”; pp. 9-34.

Yin, R.K. (1994) *Case Study Research: Design and Methods*. Sage. Beverly Hills, CA.